Language and Linguistics Compass 10/1 (2016): 14–29, 10.1111/lnc3.12174

Testing L2 Talk: A Review of Empirical Studies on Second-Language Oral Proficiency Testing

Erica Sandlund[1]*, Pia Sundqvist[1] and Lina Nyroos[2]
[1] *Karlstad University*
[2] *Uppsala University*

Abstract

In this review article, empirical studies published from 2004 through 2014 on second-language (L2) oral proficiency testing are analyzed, with a specific focus on discourse and social interaction in such tests. Taking three common test setups, oral proficiency interviews (OPIs), paired peer tests, and group peer tests as an organizing principle for the studies examined, recent developments in L2 oral proficiency testing research are situated, reviewed, and discussed, with a particular focus on tests conducted in face-to-face contexts. Findings from the review of selected journals, databases, monographs, and edited collections indicate (1) a prevalence of studies of the OPI format, but a growing research base on paired and group tests, (2) an absence of oral test studies in discourse journals, and (3) an emphasis on assessment, validation, and rater perspectives, as opposed to detailed analyses of interaction in L2 tests.

Introduction

Tests are consequential for language learners. Testing procedures are also of central interest to educational institutions, authorities, and individual teachers, as tests must be designed so that they reflect the particular aspects of a learner's language proficiency that match requirements for various placements, acceptance into education, professional positions, or national curricula. Given the increased importance given to educational testing in general, it is not surprising that research on language testing has proliferated. This review article examines empirical studies on second-language oral proficiency testing from 2004 through 2014, offering a thorough account of recent trends and developments in research on discourse and social interaction in second-language oral proficiency tests.

With the 'communicative movement' (McNamara & Roever 2006: 43) in second-language research, the focus shifted toward communicative language testing, which also entailed an interest in the joint construction of testing interaction by participants *in situ*. Such approaches take into account the 'situationally and sequentially grounded and culturally shared understandings' of speaking tests as 'particular types of speech situations' (He & Young 1998: 8). In contrast to everyday social interactions, tests are *institutional* interactions. Interactions governed by institutional constraints are, to varying degrees, goal oriented and involve particular tasks, roles, and objectives of the interaction (cf. Drew & Heritage 1992). As for test interactions, goals and tasks may be, for example, the production of assessable talk and the completion of various tasks, such as interview questions and responses, descriptive or narrative tasks. Institutional roles at play in oral proficiency tests, enacted in participants' conduct, can, for example, be 'examiner', 'teacher', or 'test-taker'. The ways in which test participants manage such specific constraints in their talk and social action, and the possible relationship between particular patterns of conduct and assessment, spurred scholarly interest in the testing context *per se*, and a number of important volumes and papers with an explicit aim to offer new tools for validating oral language tests and assessment, or for understanding the nature of social interaction in testing contexts, followed around the millennium (e.g., Brown 2003; Johnson 2000; Johnson 2001; Lazaraton 2002; Young & He 1998).

McNamara (2011: 435) argues that the 'distinctive character of language testing lies in its combination of two primary fields of expertise: applied linguistics and measurement'. The authors of this review enter the field of language testing from applied linguistics. In line with others (e.g., Ellis & Barkhuizen 2005; Gass 1997), the term *second language* (L2) is used in reference to any language – second, third, foreign, etc. – other than the learner's first language. The focus is on L2 oral proficiency testing, in particular on the *social* context of testing, that is, on discourse and social interaction in tests. Also, this review takes a specific interest in developments in the field from 2004 and onwards and aims to provide a systematic overview of recent work. Two questions guided the review:

- What is the state of the field in research on discourse and social interaction in L2 oral proficiency tests published between 2004 and 2014?
- What aspects of the testing context remain under-researched?

Although the testing of L2 oral proficiency is closely interrelated with its assessment, the latter aspect is treated peripherally (for a recent overview of L2 assessment, see Ross & Kasper 2013); here, tests are viewed as a subset of assessment, which is in line with, for example, Brown and Abeywickrama (2010). As the present review is focused on empirical studies on discourse and social interaction in L2 speaking tests, empirical studies on semi-direct tests (such as TOEFL Speaking, see, e.g., Winke & Gass 2013) or machine-controlled oral proficiency tests (such as the Versant automated oral proficiency test, see, e.g., Balogh, Barbier, Bernstein, Suzuki & Harada 2005; Downey, Farhady, Present-Thomas, Suzuki & Van Moere 2008) are beyond the scope of the review, even though such studies constitute a fairly large portion of current L2 oral proficiency testing. Conversely, several well-known tests of language proficiency do include face-to-face components and thus fall within the scope of the review, for example, the Cambridge ESOL tests (including the IELTS Speaking Test) and the Cambridge First Certificate of English (FCE), as well as various national tests of a second or foreign language. In sum, only studies with authentic face-to-face interaction as research data are included in this review.

SEARCH CRITERIA AND DELIMITATIONS

Initial searches revealed that empirical studies of oral language testing are spread across a number of publication venues, including second-language acquisition (SLA) journals, testing-focused publications, and discourse-oriented journals and edited books. Therefore, in the systematic search for empirical studies in the 2004–2014 time frame, the following three-step strategy was adopted:

(1) Considering the fact that studies of L2 oral proficiency testing are rooted in different fields, such as discourse/social interaction and applied linguistics, all issues of 15 top journals, representative of the scholarly fields concerned, were selected and trawled through. Any paper involving L2 oral proficiency testing was examined and included in the review on a relevance basis bearing the focus on discourse and social interaction in oral proficiency tests in mind (see Appendix 1 for a list of initial hits in the selected journals).

(2) For additional hits in other peer-reviewed scholarly journals, the databases *Modern Language Association/MLA*, *Education Resources Information Center/ERIC*, and *Linguistics and Language Behavior Abstracts/LLBA* were searched. Search strings applied were combinations of the following keywords: 'oral proficiency' + 'test*', 'speaking test', 'test* interaction', 'test conversation', 'OPI', and 'oral proficiency interview'. Only works published in English language peer-reviewed journals were included.

(3) In addition, library catalogues and Google Scholar were searched for monographs and edited collections on language testing. Relevant research reports from language testing organizations were also scanned for, as well as *The Language Testing and Evaluation* series. However, the review makes no claims as to having captured all publications on L2 oral proficiency testing in the given period.

In sum, the review includes original research papers published in selected peer-reviewed journals, empirical studies published in other peer-reviewed journals located through database searches, empirical studies published in edited books, and research reports that include empirical studies published in well-known book series or in research reports from various language testing organizations.

SEARCH RESULTS

As expected, the journal *Language Testing* included the highest number of hits (58). *Language Testing* and five other journals – *Applied Linguistics*, *Language Assessment Quarterly*, *Annual Review of Applied Linguistics*, *ELT Journal*, and *TESOL Quarterly* – explicitly mention 'assessment' and/or 'testing' in their aims and scope, whereas the remaining nine journals examined do not. In light of this, both *Language Assessment Quarterly* (8 hits) and *TESOL Quarterly* (7) had proportionally few hits, whereas *Foreign Language Annals* (27) had surprisingly many. The search results also reveal that seven of the journals had two or fewer hits, which probably indicate that they do not attract submissions on L2 oral testing. Although the search criteria were adhered to, decisions on which studies to include were sometimes difficult to make, mainly because of mismatches between titles and actual article content, and the number of studies with mixed methodological approaches where discursive or interactional analysis constituted a very minor part of the study as a whole.

The remainder of the present paper is structured as follows. The immediately following section situates the review in relation to developments in the field of L2 oral proficiency testing; the section also includes accounts of frequently used test formats. Next follows the review of empirical studies on discourse and social interaction in oral proficiency tests (2004–2014), divided into three subsections: candidate–examiner tests, paired tests, and group tests. These subsections, in turn, are organized on the basis of themes emerging from the studies examined.

Understanding the context: oral proficiency testing formats

L2 oral proficiency can be defined as learners' ability to converse with one or several interlocutors (cf. 'interactional competence' in an L2, Kasper & Ross 2013: 9). It is beyond the scope of this review to explore the definition in detail, but it aligns with Bachman's (1990) theoretical framework of *communicative language ability*, which encompasses five components: strategic competence, knowledge structures, language competence, psychophysiological mechanisms, and the context of situation. Bachman's work largely builds upon Canale and Swain's (1980) seminal paper, in which they make an effort to 'determine the feasibility and practicality of measuring (…) "communicative competence"' (p. 1). In their proposed theoretical framework, Canale and Swain list three main competencies: grammatical competence, sociolinguistic competence, and strategic competence. Further, they claim that the studies of sociolinguistic competence and grammatical competence are equally important to the study of communicative competence. In fact, sociolinguistic competence is said to be crucial in a theory of communicative competence and 'particularly deserving of research with respect to second language teaching and testing' (p. 17). With regard to testing, they highlight, among other things, that L2 speakers must be given the chance to participate in meaningful communicative interaction (cf. the use of the term *social interaction* in this review). With their contribution, Canale and Swain redirected scholarly attention away from a focus on test scores to one on the social construction of the assessment itself. Canale and Swain's work also underpins Bachman's (1990) concept of *language competence*, which builds upon organizational and pragmatic competence, respectively. Further, the concept of L2 oral proficiency as discussed in the present review reflects the communicative abilities as regards speaking and interaction described in the *Common European Framework of Reference for Languages* (Council of Europe 2001; see also Leclercq & Edmonds 2014) as well as the functional language ability described in the *ACTFL* (American Council on the Teaching of Foreign Languages) *Proficiency Guidelines* (ACTFL 2012).

Following upon the shift initiated in the early 1980s, in which Canale and Swain's work played a central role, a number of papers which in various ways picked up on the social construction of language testing were published. One seminal study was a paper by Van Lier (1989), which focused specifically 'what OPIs are and what the participants in them do' (489). Van Lier's study called attention to the micro-level construction of oral testing discourse and attempted to define the very nature of such social interactions in comparison to non-test conversations (see also Lazaraton 2004). Van Lier's paper was significant in directing researchers with an interest in the *process* (rather than the *outcome*) of oral proficiency testing, toward empirical studies of testing discourse and interaction.

As regards research on oral proficiency (OP) in an L2, Moeller and Theiler (2014) rightly argue that it is limited. As a consequence, research on the testing of L2 OP is even more limited and, therefore, much needed (cf. Enright 2004). Research on discourse and social interaction in OP testing contexts is also central to test construction and validation, and Lazaraton (2004: 53) states that 'it seems clear that more attention to and incorporation of discourse analysis in language test validation is needed' (see also Fulcher 1987; Shohamy 1990). However, in order to situate the review of empirical studies in the past decade, an overview of the most common formats for testing OP in face-to-face contexts follows next.

Brown and Abeywickrama (2010) suggest that language tests be classified according to their purpose and mention five test formats: achievement, diagnostic, placement, proficiency, and aptitude tests. This overview concerns placement and proficiency tests. The former attempts to place a learner at a specific level of a language curriculum or school (which may include course material), whereas the latter aims to test overall language competence.

Due to its status as a global language, English proficiency tests are very common, such as the Test of English as a Foreign Language (TOEFL) and the Cambridge ESOL (English for Speakers of Other Languages) Examinations, including the International English Language Testing System (IELTS). Upon completion of these tests, test-takers/candidates receive a certificate that indicates the level of competence. Roca-Varela and Palacios (2013) discuss the negative washback of international L2 tests: special L2 programs are designed that teach for the test only (for an overview, see Green 2013). Teaching for the test is very different from the positive washback that can be an effect of L2 OP testing (Lindblad 1992). In reference to establishing assessment criteria in L2 tests, then, Roca-Varela and Palacios (2013: 66) suggest that 'oral tests should become more human'. The three proficiency tests listed above all include a module for speaking, but there are also pure L2 speaking tests.

CANDIDATE-EXAMINER TESTS

L2 speaking/OP tests date back to the early 20th century, but it was not until during the second world war that such tests shifted from having had a focus on dictation and pronunciation to a focus on the ability to perform/interact (Fulcher 2003). The three-part test format used then, picture description, sustained speech, and conversation (for details, see Lado 1961), was a precursor of the first published speaking test, the Foreign Service Institute (FSI) Oral Proficiency Interview, known as the OPI (Fulcher 2003). The three-part design of OP tests is still frequently used – lower-level abilities being tested first – such as moving from monologue via dialogue to interaction in paired tests (cf. Sandlund & Sundqvist 2011). The OPI (also referred to as Language Proficiency Interview, LPI, see Young & He 1998) is a standardized, global assessment of OP and involves an examiner and a test-taker (audio-recorded). The candidate's output is assessed by the examiner and a second rater against pre-set criteria for 10 levels, described in the aforementioned ACTFL guidelines (ACTFL 2012).

Similar to the English tests mentioned above, the OPI is often required for various reasons, for example, for teacher candidates (Burke 2013; Glisan, Swender & Surface 2013; Kissau 2014) and study-abroad programs (see, e.g., Di Silvio, Donovan & Malone 2014; Diao, Freed & Smith 2011; Golonka 2006; Lindseth 2010; Magnan & Back 2007; Segalowitz & Freed 2004). There are also so-called simulated OPIs (SOPIs) (see, e.g., Kenyon 2000), in which the test-taker is given computer- or tape-mediated prompts along with a printed test booklet and responds by speaking into a microphone (Kenyon & Malabonga 2001; for a review of research comparing OPIs and SOPIs, see Stansfield & Kenyon 1992). Last, the OPI has been criticized for being behavioristic (Johnson 2000), for the way the examiner is implicated in the construction of test-takers' output (Brown 2003) and for not resembling natural conversation due to the asymmetric relationship between the examiner and the test-taker (Green 2014; Van Lier 1989). In addition, findings in Winke and Gass (2013) suggest that when examiners/raters are familiar with the test-takers' first language, they tend to orient to the learners' talk in a biased way, which compromises test reliability; moreover, different kinds of preparation for OPI affect test scores (Sullivan 2011). Others claim the OPI yields a valuable and independent measure of OP (Cubillos 2010) resembling real-life speaking skills (Kenyon & Malabonga 2001).

PAIRED TESTS

In contrast to OPIs, there are paired/dyadic L2 speaking tests, which have grown increasingly popular (Nitta & Nakatsuhara 2014). Presumably, paired tests better resemble natural conversation (Ducasse & Brown 2009). A study by Brooks (2009: 341) compared the two test formats, OPI and paired, and found that the latter resulted not only in higher OP scores for the test-takers but also in 'more interaction, negotiation of meaning, consideration of the interlocutor and more complex output'. In a similar (but theoretical) comparison, Birjandi (2011) concludes that paired testing might indeed be more beneficial than traditional OPIs but calls for more empirical research on the topic. Whatever the case, paired tests are clearly problematic from an assessment perspective since the spoken output is a joint product (He & Young 1998) by two test-takers later assessed individually (May 2011b; Sandlund & Sundqvist 2011). Moreover, the role of the interlocutor is crucial since s/he is likely to influence both scores – sometimes positively, sometimes negatively (Davis 2009; Galaczi 2008; Iwashita 2001) – and interaction (Lazaraton & Davis 2008).

There are also questions as regards possible effects of pre-task planning on performance; whereas a previous study shows a positive relationship between planning time and OP scoring (Mehnert 1998), a more recent one reveals limited effects of planning on output and also that planning may strip test-takers of the opportunity to demonstrate their interactional abilities (Nitta & Nakatsuhara 2014; see also Wigglesworth & Elder 2010). Regardless of test format, type of preparation affects test results (Farnsworth 2013; Huang & Hung 2013; Issitt 2008), and it appears that gender also may play a role in oral testing (Amjadian & Ebadi 2011; McKay 2006; Reemann, Alas & Liiv 2013).

GROUP TESTS

Based on the results from our searches, studies on paired or group tests are much less common than studies on tests involving one test-taker. As regards group tests, generally three or four candidates participate, and the examiner remains silent. Tasks and prompts vary, but the point is to elicit group interaction (and it happens that the examiner is supposed to participate as well, Green 2014).

A frequently used group test of L2 English OP was developed in Japan (Bonk & Ockey 2003): candidates are given a short written prompt about a general topic and after 1 min of preparation, they begin a discussion that should last approximately 10 min (Bonk & Ockey 2003; Leaper & Riazi 2014). In other tests, topic cards (see, e.g., Sandlund & Sundqvist 2013) or pictures (see, e.g., Hasselgren 2000) are used to elicit discussion. Among other things, recent studies on L2 group tests have treated topic negotiation, turn-taking practices, and more, which is developed further below.

Empirical studies of discourse and social interaction in L2 OP tests 2004–2014

In this section, empirical studies of discourse and social interaction in L2 oral proficiency tests published in 2004–2014 are discussed. Given the variety in analytical approaches and mixed-methods setups in studies that involved authentic face-to-face discourse data, a separation of studies identifying as 'discourse analytic' and 'social interactional' respectively was not possible to make, since many of the studies reviewed combined different methodological approaches or were rather brief in their descriptions of analytical procedures. Therefore, the review below utilizes the different test setups as an organizing principle and, as a consequence, addresses both discourse-oriented methods and conversation analytic studies.

DISCOURSE AND SOCIAL INTERACTION IN CANDIDATE-EXAMINER TESTS

Even though OPIs always involve an examiner, generally a native speaker, the interview setup varies and may involve role-plays, interview questions, or particular tasks to be accomplished. Research on discourse and/or social interaction in OPIs also targets such differences but generally treats OPIs as a particular form of social interaction that affords particular interactional behavior. In the 2004–2014 time frame, four recurrent topical areas were identified: (i) the oral proficiency interview (OPI) as a unique type of social event, (ii) the OPI compared to other test constructs/conditions, such as role-plays or paired formats, (iii) OPI examiner conduct and inter-interviewer variation, and (iv) specific features of OPI interaction, such as question design, responses, locally constructed identities, or politeness.

The OPI as a Unique Type of Social Event

In the first topical theme identified, efforts to define the type of social event that OPIs constitute are made. Given the aim of eliciting candidates' L2 conversational competence, many studies focus on defining the type of social interaction that such interviews may represent, that is, whether OPIs resemble ordinary conversations and whether this interactional type of event is suitable for assessing L2 communicative skills (Simpson 2006) (cf. also older extensive reviews in Johnson & Tyler 1998; Lazaraton 2002; Young & He 1998). A number of studies in our review adhere to conversation analysis (CA), an ethnomethodologically grounded approach to the study of the organization of social interaction (Sacks, Schegloff & Jefferson, 1974; Sidnell & Stivers, 2013). From a CA perspective, Seedhouse (2013) argues that OPIs constitute a particular variety of interaction different from, for example, L2 classroom interaction. Basing his analysis on data from the IELTS Speaking Test (IST, see also Seedhouse & Egbert 2006), Seedhouse compares patterns of turn-taking, repair, and the relationship between institutional goals and participant orientations in L2 classrooms and university interactions and shows that patterns of repair differ significantly between the OPIs and the other two settings (cf. also Egbert 1998).

Aside from the strict turn allocation system and interactional asymmetry (candidate-examiner), the high demands on standardization only allows for particular forms of repair initiated by the examiner. Moreover, a striking feature of the IST is that there is 'no requirement to achieve intersubjectivity', that is, shared understanding based on negotiation of the meaning of turns and actions (Seedhouse 2013: 211), as would be the case in other interactional varieties. Among other things, Seedhouse argues that findings from his study may inform test construction and rating scales.

In a similar study, van Compernolle (2011) examined LPIs between a teacher and L2 French learners, focusing on precision timing and the conditional relevance of responses to questions, repairs, and nonunderstanding of questions. The sequential pattern of LPI talk identified was much like classroom interaction (teacher's question, student's response, and teacher's closing third turn, such as 'okay'). It is argued that responses to LPI questions index the learner's interactional competence *as interviewees* and that the preference of organization for providing a response overrides the appropriateness of the content of the response. The specific interactional competence that students can demonstrate relates to their 'socially constructed knowledge of what it means to interact with an LPI interviewer' and the 'rhetorical script' of the interview being enacted (p. 132), an observation that causes van Compernolle to join in existing critique of the interview format for assessing L2 OP.

The OPI Compared to other Test Conditions

In a second set of studies, the OPI is compared to other test constructs/conditions, such as role-plays (e.g., Halleck 2007; Hüttner 2014; Lorenzo-Dus & Meara 2004) or paired formats (Brooks 2009). Halleck's (2007) findings result in a critique of dialogic role-play tasks as valid measures of a candidate's oral proficiency because of the dominant role in constructing the test discourse played by the examiner. Similarly, Okada (2010) investigates role-plays in high-stake OPIs, concluding that role-plays are not necessarily a better option as compared to OPIs in the sense that the interviewer to a large extent organizes turn-taking as well as topic initiation. However, Okada's study also demonstrates some positive aspects of role-plays; they may offer opportunities for test-takers to display a variety of interactional competencies as well as different discursive identities.

OPI Examiner Conduct and Inter-interviewer Variation

Not surprisingly, a dominant theme for OPI interaction research is examiner conduct and inter-interviewer variation (Brown 2005; Kondo-Brown 2004; Lorenzo-Dus & Meara 2005; Nakatsuhara 2008; Plough & Bogart 2008; Reemann et al. 2013; Tominaga 2013). A number of studies prior to our selected time frame suggest that variations in interviewer strategies play a central role for candidate performance, which is confirmed by Nakatsuhara (2008), who observed that two interviewers who talked to the same candidate exhibited different patterns of questioning, in particular when a question was 'unsuccessful' (270) in eliciting a response. She concludes that the candidate scored better on fluency and pronunciation with the examiner who used a less structured interview style, more so-called receipt tokens, and positive assessments. Similarly, Ross (2007) identified differences in interviewer styles in two OPIs with the same candidate, which affected the candidate's score.

Kasper and Ross (2007) investigate interviewer questions. They argue that test-takers are given different conditions regarding topic understanding depending on the interviewer's inclination to pose multiple questions in a single turn or if questions are delivered 'one at time' in a question–answer fashion. In other words, interviewers' question design might 'inject bias into the interview process' (2067). The authors recommend a shift in focus in interviewer training from questions as tools for eliciting assessable output to questions as central to candidates' topic understanding and conclude that differences in question design might be crucial for the validity of the OPI (cf. also Nyroos & Sandlund 2014). Closely related to issues of validity are papers that touch upon test standardization, but which do not focus solely on the interviewer's conduct. By way of example, Norton (2013) shows how interviewer and test-taker locally construct situated identities, displayed in their orientation to different discursive frames, which in turn may compromise standardization.

Specific Features of OPI Interaction

The fourth category identified shows the greatest variation, but a common denominator is an interest in specific features of the structure and organization of OPIs, such as interviewers' pursuit of a relevant response in role-play OPIs (Okada & Greer 2013), the longitudinal development of candidates' ability to do storytelling in extended turns (Tominaga 2013), interviewers' orientation to politeness in OPI request sequences (Kasper 2006), or repair and task management (Kasper 2013). Almost all studies falling under this theme are conversation analytic, which is not surprising given the core CA enterprise of studying structural organization of talk in turns and sequences of authentic interaction.

Demonstrated through a detailed analysis of particular aspects of discourse and social interaction in OPIs, many of the studies on the OPI as a social event reviewed above align with earlier critique of the OPI as a holistic measure of a test-taker's L2 oral proficiency. Although most of the studies point to problems with the OPI format associated with asymmetry between participants, variations across interviews, and the type of competence that candidates *can* display in an interview format, researchers behind them tend to recommend that their findings inform test construction and rater training. Now turning to discourse and social interaction in paired OP tests, some of the themes from OPI research reoccur, such as the comparison between the paired format and others; however, as demonstrated below, themes unique to the paired format, such as the interaction between peers of different proficiency levels, surface.

DISCOURSE AND SOCIAL INTERACTION IN THE PAIRED FORMAT

Studies focusing on interaction in paired OP tests increased toward the end of our time period. A few studies compare OPIs with the paired format (e.g., Brooks 2009; Hüttner 2014), and a general finding of these studies – other than that the test format affects test performance – seems to be that the paired format allows for more flexible candidate contributions and a wider range of more complex actions and negotiations of meaning. As mentioned, although it has been argued that peer–peer interaction resembles natural conversation, the challenge with the paired format has to do with assessment: individual scoring of a jointly accomplished 'product' (the test interaction) is problematic. In light of this problem, Ducasse (2010: 1) argues that

> not enough is known about what takes place in the interaction between students in a paired oral proficiency test, and as a consequence, no rating scales have been developed based directly on empirical data from observed performances of such interactions.

Her study included rater and candidate perceptions of salient features of successful peer interaction (cf. also May 2009, 2011a, 2011b) and the development of a rating scale based on recordings and analysis of peer interaction. The study is mainly focused on how features such as 'interactive listening', 'non-verbal communication', and 'interactional management' were salient to both raters and candidates viewing their test partner's performance and on how these features can be incorporated into rating procedures. Despite the repeated mention of peer interaction, the study does not rely upon analysis of the test discourse in itself (cf. also Ducasse & Brown 2009).

Paired Test Interaction and Assessment

The interest in rating scales and the connections between assessment and interaction is a prevalent theme among our hits. An often cited article is Galaczi (2008), involving the First Certificate in English. Using a discourse analytic approach informed by CA, along with quantitative methods, she identified three patterns of how test-takers orient to each other, which all signalled varying degrees of cooperation, mutuality, and equality. In terms of scores, the 'collaborative pattern' was deemed to be the most successful. Galaczi suggests that candidates with limited L2 skills are not particularly engaged in interactions with their interlocutors, a track explored further in Galaczi (2014), where she recommends a broader view on interactional competence. Along similar lines, Butler and Zeng (2014) found that fourth-graders used more formulaic/fixed turn allocation strategies than sixth-graders, concluding that turn allocation is an advanced interactional competence.

Other studies that deal with the relationship between interaction and assessment in the paired format include Sandlund & Sundqvist (2011), who analyze sequences from the national test of English as a foreign language for ninth-graders in Sweden. The study adopts CA and compares the sequential analysis with assessment data, focusing on how students manage interactional trouble connected to the test tasks ('task-related trouble'). The findings indicate that certain task management strategies appeared to be rated less favorably than others, even though some of them (such as task abandonment and negotiation of understanding of the test task) were perfectly productive for the students themselves in terms of test-wiseness, a concept also discussed in work on language testing in general (see, e.g., Bachman 1990; Brown & Abeywickrama 2010; Mousavi 2012).

The Interlocutor

Yet other studies pay attention to the effect of the interlocutor on test performance, a topic generally explored through quantitative studies (e.g., Bennett 2012; Csépes 2009). In Norton's (2005) study on the Cambridge Speaking Test, the author argues for a connection between pairing and performance if the candidates are at different proficiency levels and that gender and familiarity with the other candidate potentially plays a role for linguistic performance (but see Lazaraton's 2006 critique of the reliability of the findings of Norton's analyses). A radically different approach to analyzing the effect of pairing is Lazaraton and Davis's (2008) CA study of proficiency identities. The authors show that the candidates bring proficiency identities to the test and that these might be ratified, maintained, or modified depending on the interlocutor. This particular study demonstrates how proficiency should be considered a skill that test-takers dynamically *construct in* peer–peer interaction, rather than something only being *displayed in* interaction.

Test Tasks

Just as the general test setup has been shown to affect interactional organization, so do individual test tasks. In a recent study (Sandlund & Sundqvist 2013), a CA approach to task understandings in paired L2 tests is presented. The study offers a detailed analysis of how a particular discussion topic (presented on a card) is demonstrably treated differently by test-takers and the teacher/examiner respectively during the test. While test-takers orient to producing 'any' kind of contributions to the topic, the teacher displays a more restricted understanding of the task, evident in her repeated attempts to steer the interaction back to the verbatim formulation on the topic card. The study underscores the contribution of microanalytic studies of task understandings to test construction/assessment purposes and also points to the need for more discussion on the relative importance of task adherence in assessing OP.

INTERACTION IN GROUP OP TESTS

The third format of OP testing is small groups, which is a growing area in the research reviewed. This is not surprising given the relative novelty of group testing. Two central research interests surface: (i) the unique features of group interaction in L2 testing contexts and (ii) the effect of various interlocutor variables on test discourse and performance.

Unique Features of Group Testing Interaction

In the first set of studies, Gan, Davison, and Hamp-Lyons (2009) examine how topics are introduced and negotiated. They observe that peer-group discussions force the candidates to relate to each other in the ongoing interaction, to monitor the ongoing talk, and to identify the assessment task agenda. In other words, group settings might provide test-takers with opportunities to demonstrate 'real-life' interactional abilities. The additional complexity of having several participants to relate to in a testing context is also addressed by Greer and Potter (2012), who point to the challenges of turn-taking in multi-party interaction. The authors argue that turn-taking management in group interaction should be viewed as an indicator of interactional competence. In a related study, Greer and Potter (2008) examine peer-driven questions, showing that proficient speakers take the role of assigning speakership also to reticent participants.

Interlocutor Effects on Test Discourse and Performance

In the second group of studies, several studies examine test-takers' varying proficiency levels to group test interactional patterns (e.g., Gan 2010; Nakatsuhara 2011). Gan compared higher- and lower-scoring test-takers' group interactions. The lower-scoring students oriented to the pre-set test prompts to a greater extent than the higher-scoring students (a similar pattern observed by Sandlund & Sundqvist 2011), but both groups displayed a wide range of interactional skills. Gan (2010: 599) concludes that the group format 'authentically reflects candidates' interactional skills and their moment-by-moment construction of social and linguistic identity'. Finally, Luk (2010: 49) discusses how test-takers 'stage a performance' in order to come across as competent interlocutors in group tests, thereby revealing a degree of inauthenticity in their conversation, and recommends more research into impression management in OP tests.

Conclusion

This review article has examined empirical studies on L2 oral proficiency testing published between 2004 and 2014 with a particular focus on studies on discourse and social interaction in such tests. Interestingly, a majority of the studies examined did not include discourse data at all, which might be a reflection on authorship; that is, whether authors come from the field of measurement or applied linguistics (cf. McNamara 2011). Searches also revealed that studies on OPIs were much more frequent than studies on paired or group tests. There was an increase of paired/group studies over the last few years of our set time frame, possibly mirroring the social turn within the broad field of second-language acquisition research (see Ortega 2011).

A number of hits in the database searches had key terms, such as *interaction and oral proficiency*, in the paper titles, but a closer look revealed that many of them did not, in fact, include any reproduced segments of conversation (e.g., Ducasse 2010; May 2011a; Sato & Lyster 2012). Thus, they did not set out to provide knowledge *on* interaction, but rather *about* interaction.

Another bulk of studies did use discourse data, but mainly for validating rating scales or developing descriptions of interactional patterns to be compared to ratings (Galaczi 2008). As a result, the number of hits that turned out to be directly relevant to the review decreased greatly. In fact, relatively few studies (2004–2014) explored specific features or the sequential development of L2 testing interaction. However, from those that do, it is evident that interaction patterns vary a lot between test conditions as they set up very different *frames* (cf. Goffman 1974). Frames for the testing context, then, would represent different ways of structuring and interpreting the social context in which an individual participates. In the context of L2 OP testing, being tested in a structured interview or in a more peer-driven conversation format presents different challenges for the test-taker depending on the conceptions inherent in the test frame. OPIs leave little maneuver space for test-takers as the interviewer is in charge of the sequential development, which in turn discursively defines the role of an interviewer (cf. e.g., Sacks 1992). For this reason, raised awareness regarding the specific constraints permeating the interaction in OPIs is desirable (cf. Johnson 2001). When L2 oral proficiency is the construct, the OPI is an unsuitable context since candidates' interactional space is restricted in terms of actions, topic initiation, and turn-taking management. The OPI interaction pattern can be compared to a circle, where the interviewer sets the course by means of the question type, which is being responded to by the candidate. When the interviewer is satisfied with a response, a new circle is initiated, and the test-taker is locked into a discursive frame. Consequently, some advocate role-play tasks in OPIs in which discursive roles are less fixed, leaving room for a wider range of conversational practices. It is safe to say that the more conversation-like the context, the more dynamic the interaction, leading to more opportunities for the candidates to demonstrate interactional competence. Instead, problems for the standardization of test procedures arise.

Finally, we would like to comment briefly on some methodological aspects of the studies examined here. It was noticeable that descriptions of data and analytic approaches at times were conspicuously brief or lacked relevant information, for example, on test-takers' first language. Further, while a number of studies claim an interest in interactional patterns, relatively few conduct the type of sequential analysis characteristic of CA, and some of those that claim a CA framework do not adhere to the emic focus that distinguishes CA from other discourse-oriented approaches (see, for instance, Okada's (2010) critique of Kormos (1999) and Roever's (2011) comments on Walters (2007, 2009)). The fact that CA may productively inform the design of L2 assessment tasks and offer insights into particular testing formats, for example role-plays and group discussions, has been argued by Schegloff, Koshik, Jacoby, and Olsher (2002). A final observation, then, is that interdisciplinary collaboration as well as cross-fertilization between different discursive approaches may bring current knowledge of L2 oral testing forward.

Acknowledgement

This research was supported by the Swedish Research Council, grant no. 2012-4129.

Short Biographies

Erica Sandlund is an Associate Professor in English Linguistics. She has an educational background in English linguistics and psychology and received her PhD degree from Karlstad University (2004). Her doctoral dissertation *Feeling by Doing: The social organization of everyday emotions in academic talk-in-interaction* (Karlstad University Studies, 2004:36) was a conversation analytic exploration of displays of affect in social interaction. She has continued to work in the field of conversation analysis with a primary interest in institutional settings, such as academic seminars, performance appraisal interviews, classrooms, and second-language speaking tests. Analytic interests include affect displays, task management, question design, code switching, and reported speech enactments. Through her work on performance appraisal interviews, she has been affiliated with the Center for Gender Studies at Karlstad University and has taught several graduate courses on language, gender, and interaction and co-edited a special issue of *Economic and Industrial Democracy* on gender and sustainable regional development. Currently, she coordinates a 4-year research project on the speaking part of the national test of English in Sweden (Swedish Research Council Reg. no 2012-4129), *Testing Talk*, in which interaction in and assessment of L2 oral proficiency tests is the focus. Together with her co-authors, she has published a number of papers on paired test interaction and on performance appraisal interviews as an arena for the management of organizational norms. As a graduate student, she was a visiting scholar at the School of Communication, San Diego State University.

Pia Sundqvist is an Associate Professor in English Linguistics. She received her PhD degree from Karlstad University (2009) with the doctoral dissertation *Extramural English Matters: Out-of-School English and Its Impact on Swedish Ninth Graders' Oral Proficiency and Vocabulary*. Her research is focused on second-language acquisition, in particular computer-assisted language learning (CALL), and education. She has authored papers in these areas for journals such as *ReCALL, Novitas-ROYAL, The European Journal of Applied Linguistics and TEFL*, and *Apples*. Furthermore, she has contributed with book chapters in recent publications in the field of informal learning of languages in out-of-school contexts: *Beyond the Language Classroom* (edited by Benson and Reinders, Palgrave Macmillan 2011) and *Digital Games in Language Learning and Teaching* (edited by Reinders, Palgrave Macmillan 2012). She is currently involved in two projects funded by the Swedish Research Council, Testing Talk (see above) and *Bridging the Gap between In- and Out-of-School English* (Swedish Research Council Reg. no. 2013-785). Sundqvist has extensive experience from teaching English, Swedish, and Spanish in secondary and upper-secondary school. As a graduate student, she studied applied English linguistics at the University of Houston, Texas.

Lina Nyroos has a PhD degree from the Department of Scandinavian Languages at Uppsala University, Sweden. Her dissertation (2012), *The Social Organization of Institutional Norms: Interactional Management of Knowledge, Entitlement and Stance*, focused on institutional traits in interaction in two different settings: group tutoring sessions at university level and performance appraisal interviews in organizations, using conversation analysis. She currently holds a postdoctoral position at Uppsala University and is one of the researchers in a 4-year project on the speaking part of the national test of English in Sweden (Swedish Research Council Reg. no. 2012-4129), *Testing Talk*, which targets interaction in and assessment of L2 oral proficiency tests. She is also a member of a Swedish think tank for the humanities, HumTank, which aims to promote, debate, and make visible research in the humanities. In a project on parliamentary discourse in the European Union, she has studied interaction patterns in parliamentary debates. Among her research interests are questions and their design in social interaction, and together with Sandlund, she recently published a paper on the interactional management of standardized question items in performance appraisal interviews (*Pragmatics & Society* 5:2, 2014). From 2016, she is the Director of Studies at the Department of Scandinavian Languages, Uppsala University.

Note

* Correspondence address: Erica Sandlund, Faculty of Arts and Social Sciences, Karlstad University, SE-65188 Karlstad, Sweden. E-mail: erica.sandlund@kau.se

Works Cited

ACTFL. 2012. *ACTFL proficiency guidelines*. Alexandria, VA: ACTFL.

Amjadian, Mohiadin, and Saman Ebadi. 2011. Variationist perspective on the role of social variables of gender and familiarity in L2 learners' oral interviews. *Theory and Practice in Language Studies* 1. 722–28.

Bachman, Lyle F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Balogh, Jennifer E., Isabella Barbier, Jared Bernstein, Masanori Suzuki, and Yasunari Harada. 2005. A common framework for developing automated spoken language tests in multiple languages. *Japanese Journal for Research on Testing (JART)* 1. 67–79.

Bennett, Rita. 2012. Is linguistic ability variation in paired oral language testing problematic? *ELT Journal* 66. 337–46.

Birjandi, Parviz 2011. From face-to-face to paired oral proficiency interviews: the nut is yet to be cracked. *English Language Teaching* 4. 169–75.

Bonk, William J., and Gary J. Ockey. 2003. A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing* 20. 89–110.

Brooks, Lindsay. 2009. Interacting in pairs in a test of oral proficiency: co-constructing a better performance. *Language Testing* 26. 341–66.

Brown, Annie. 2003. Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20. 1–25.

——. 2005. *Interviewer variability in oral proficiency interviews*. Frankfurt am Main: Peter Lang.

Brown, H. Douglas, and Priyanvada Abeywickrama. 2010. *Language assessment. Principles and classroom practices*. White Plains, NY: Pearson Education.

Burke, Brigid M. 2013. Looking into a crystal ball: is requiring high-stakes language proficiency tests really going to improve world language education? *The Modern Language Journal* 97. 531–34.

Butler, Yuko Goto, and Wei Zeng. 2014. Young foreign language learners' interactions during task-based paired assessments. *Language Assessment Quarterly* 11. 45–75.

Canale, Michael and Merrill, Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1. 1–47.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Csépes, Ildikó. 2009. *Measuring oral proficiency through paired-task performance*. Frankfurt am Main: Peter Lang.

Cubillos, Jorge. 2010. Computer-mediated oral proficiency assessments: validity, reliability and washback. *International Journal of Technology, Knowledge and Society* 6. 85–102.

Davis, Larry. 2009. The influence of interlocutor proficiency in a paired oral assessment. *Language Testing* 26. 367–96.

Di Silvio, Francesca, Anne Donovan, and Margaret E. Malone. 2014. The effect of study abroad homestay placements: participant perspectives and oral proficiency gains. *Foreign Language Annals* 47. 168–88.

Diao, Wenhao, Barbara Freed, and Leigh Smith. 2011. Confirmed beliefs or false assumptions? A study of home stay experiences in the French study abroad context. *Frontiers: The Interdisciplinary Journal of Study Abroad* 21. 109–42.

Downey, Ryan, Hossein Farhady, Rebecca Present-Thomas, Masanori Suzuki, and Alistair Van Moere. 2008. Evaluation of the usefulness of the Versant for English test: a response. *Language Assessment Quarterly* 5. 160–67.

Drew, Paul, and John Heritage (eds). 1992. *Talk at work. Interaction in institutional settings*. Cambridge: Cambridge University Press.

Ducasse, Ana Maria. 2010. *Interaction in paired oral proficiency assessment in Spanish*. Frankfurt am Main: Peter Lang.

Ducasse, Ana Maria, and Annie Brown. 2009. Assessing paired orals: raters' orientation to interaction. *Language Testing* 26. 423–43.

Egbert, Maria. 1998. Miscommunication in language proficiency interviews of first-year German students: a comparison with natural conversation. Talking and testing. *Discourse approaches to the assessment of oral proficiency*, ed. by Richard Young, and Agnes Weiyun He, 147–69. Amsterdam: John Benjamins.

Ellis, Rod, and Gary Barkhuizen. 2005. *Analysing learner language*. Oxford: Oxford University Press.

Enright, Mary K. 2004. Research issues in high-stakes communicative language testing: Reflections on TOEFL's new directions. *TESOL Quarterly* 38. 147–51.

Farnsworth, Tim. 2013. Effects of targeted test preparation on scores of two tests of oral English as a second language. *TESOL Quarterly* 47. 148–55.

Fulcher, Glenn. 1987. Tests of oral performance: the need for data-based criteria. *ELT Journal* 41. 287–91.

——. 2003. *Testing second language speaking*. Harlow: Pearson Education.

Galaczi, Evelina D. 2008. Peer–peer interaction in a speaking test: the case of the First Certificate in English examination. *Language Assessment Quarterly* 5. 89–119.

—— 2014. Interactional competence across proficiency levels: how do learners manage interaction in paired speaking tests? *Applied Linguistics* 35. 553–74.

Gan, Zhengdong. 2010. Interaction in group oral assessment: a case study of higher- and lower-scoring students. *Language Testing* 27. 585–602.

Gan, Zhengdong, Chris Davison, and Liz Hamp-Lyons. 2009. Topic negotiation in peer group oral assessment situations: a conversation analytic approach. *Applied Linguistics* 30. 315–44.

Gass, Susan. 1997. *Input, interaction, and the second language learner*. Mahwah, NJ: Lawrence Erlbaum.

Glisan, Eileen W., Elvira Swender, and Eric A. Surface. 2013. Oral proficiency standards and foreign language teacher candidates: current findings and future research directions. *Foreign Language Annals* 46. 264–89.

Goffman, Erving. 1974. *Frame analysis*. New York, NY: Harper and Row.

Golonka, Ewa M. 2006. Predictors revised: Linguistic knowledge and metalinguistic awareness in second language gain in Russian. *The Modern Language Journal* 90. 496–505.

Green, Anthony. 2013. Washback in language assessment. *International Journal of English Studies* 13. 39–51.

——. 2014. *Exploring language assessment and testing: language in action*. Oxon: Routledge.

Greer, Tim, and Hitomi Potter. 2008. Turn-taking practices in multi-party EFL oral proficiency tests. *Journal of Applied Linguistics* 5. 297–320.

Greer, T., and H. Potter. 2012. Turn-taking practices in multi-party EFL oral proficiency tests. *Journal of Applied Linguistics* 5. 297–320.

Halleck, Gene. 2007. Data generation through role-play: assessing oral proficiency. *Simulation & Gaming* 38. 91–106.

Hasselgren, Angela. 2000. The assessment of the English ability of young learners in Norwegian schools: an innovative approach. *Language Testing* 17. 261–77.

He, Agnes Weiyun, and Richard Young. 1998. Language proficiency interviews: a discourse approach. *Talking and testing: discourse approaches to the assessment of oral proficiency*, ed. by Richard Young, and Agnes Weiyun He, 1–24. Amsterdam: John Benjamins.

Huang, Heng-Tsung Danny, and Shao-Ting Alan Hung. 2013. Comparing the effects of test anxiety on independent and integrated speaking test performance. *TESOL Quarterly* 47. 244–69.

Hüttner, Julia. 2014. Agreeing to disagree: 'doing disagreement' in assessed oral L2 interactions. *Classroom Discourse* 5. 194–215.

Issitt, Steve. 2008. Improving scores on the IELTS speaking test. *ELT Journal* 62. 131–38.

Iwashita, Noriko. 2001. The effect of learner proficiency on interactional moves and modified output in nonnative–nonnative interaction in Japanese as a foreign language. *System* 29. 267–87.

Johnson, Marysia. 2000. Interaction in the oral proficiency interview: problems of validity. *Pragmatics* 10. 215–31.

——. 2001. *The art of nonconversation: a reexamination of the validity of the oral proficiency interview*. Baltimore, MD: Yale University Press.

Johnson, Marysia, and Andrea Tyler. 1998. *Re-analyzing the OPI: how much does it look like natural conversation? Talking and testing: discourse approaches to the assessment of oral proficiency*, ed. by Richard Young, and Agnes Weiyun He, 27–51. Amsterdam: John Benjamins.

Kasper, Gabriele. 2006. When once is not enough: politeness of multiple requests in Oral Proficiency Interviews. *Multilingua* 25. 323–50.

——. 2013. Managing task uptake in oral proficiency interviews. *Assessing second language pragmatics*, ed. by Steven J. Ross and Gabriele Kasper, 258–87. Basingstoke: Palgrave Macmillan.

Kasper, Gabriele, and Steven J. Ross. 2007. Multiple questions in oral proficiency interviews. *Journal of Pragmatics* 39. 2045–70.

——. 2013. Assessing second language pragmatics: an overview and introductions. *Assessing second language pragmatics*, ed. by Steven J. Ross, and Gabriele Kasper, 1–40. Bristol: Palgrave Macmillan.

Kenyon, Dorry Mann. 2000. Tape-mediated oral proficiency testing: considerations in developing Simulated Oral Proficiency Interviews (SOPIs). *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests*, ed. by S. Bolton, 87–106. München: Goethe-Institut.

Kenyon, Dorry Mann, and Valerie Malabonga. 2001. Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language, Learning and Technology* 5. 60–83.

Kissau, Scott. 2014. The Impact of the Oral Proficiency Interview on one foreign language teacher education program. *Foreign Language Annals* 47. 527–45.

Kondo-Brown, Kimi. 2004. Investigating interviewer–candidate interactions during oral interviews for child L2 learners. *Foreign Language Annals* 37. 601–13.

Kormos, Judit. 1999. Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing* 16. 163–88.

Lado, Robert. 1961. *Language testing: the construction and use of foreign language tests*. London: Longman.

Lazaraton, Anne. 2002. *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

——. 2004. Qualitative research methods in language test development and validation. European language testing in a global context. *Proceedings of the ALTE Barcelona Conference July 2001*, ed. by Michael Milanovic, and Cyril Weir, 51–73. Cambridge: Cambridge University Press.

——. 2006. Process and outcome in paired oral assessment. *ELT Journal* 60. 287–89.

Lazaraton, Anne, and Larry Davis. 2008. A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly* 4. 313–35.

Leaper, David A., and Mehdi Riazi. 2014. The influence of prompt on group oral tests. *Language Testing* 31. 177–204.

Leclercq, Pascale, and Amanda Edmonds. 2014. How to assess L2 proficiency? An overview of proficiency assessment research. *Measuring L2 proficiency: Perspectives from SLA*, ed. by Pascale Leclercq, Amanda Edmonds, and Heather Hilton, 3–23. Bristol: Multilingual Matters.

Lindblad, Torsten. 1992. Oral tests in Swedish schools: a five-year experiment. *System* 20. 279–92.

Lindseth, Martina U. 2010. The development of oral proficiency during a semester in Germany. *Foreign Language Annals* 43. 246–68.

Lorenzo-Dus, Nuria, and Paul Meara. 2004. Role-plays and the assessment of oral proficiency in Spanish. *Current trends in the pragmatics of Spanish*, ed. by Rosina Márquez Reiter, and María Elena Placencia, 79–97. Philadelphia, PA: John Benjamins.

——. 2005. Examiner support strategies and test-taker vocabulary. *IRAL* 43. 239–58.

Luk, Jasmine. 2010. Talking to score: impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly* 7. 25–53.

Magnan, Sally Sieloff, and Michele Back. 2007. Social interaction and linguistic gain during study abroad. *Foreign Language Annals* 40. 43–61.

May, Lyn. 2009. Co-constructed interaction in a paired speaking test: the rater's perspective. *Language Testing* 26. 397–421.

——. 2011a. *Interaction in a paired speaking test*. Frankfurt am Main: Peter Lang.

——. 2011b. Interactional competence in a paired speaking test: features salient to raters. *Language Assessment Quarterly* 8. 127–45.

McKay, Penny. 2006. *Assessing young language learners*. Cambridge: Cambridge University Press.

McNamara, Tim. 2011. Applied linguistics and measurement: a dialogue. *Language Testing* 28. 435–40.

McNamara, Tim, and Carsten Roever. 2006. *Language testing: the social dimension*. Oxford: Blackwell.

Mehnert, Uta. 1998. The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition* 20. 83–107.

Moeller, Aleidine J., and Janine Theiler. 2014. Spoken Spanish language development at the high school level: a mixed-methods study. *Foreign Language Annals* 47. 210–40.

Mousavi, Seyyed Abbas. 2012. *An encyclopedic dictionary of language testing*. Tehran: Rahnama Press.

Nakatsuhara, Fumiyo. 2008. Inter-interviewer variation in oral interview tests. *ELT Journal* 62. 266–75.

——. 2011. Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing* 28. 483–508.

Nitta, Ryo, and Fumiyo Nakatsuhara. 2014. A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing* 31. 147–75.

Norton, Julie. 2005. The paired format in the Cambridge Speaking Tests. *ELT Journal* 59. 287–97.

——. 2013. Performing identities in speaking tests: co-construction revisited. *Language Assessment Quarterly* 10. 309–30.

Nyroos, Lina, and Erica Sandlund. 2014. From paper to practice: Asking and responding to a standardized question item in performance appraisal interviews. *Pragmatics & Society* 5. 165–190.

Okada, Yusuke. 2010. Role-play in oral proficiency interviews: interactive footing and interactional competencies. *Journal of Pragmatics* 42. 1647–68.

Okada, Yusuke, and Tim Greer. 2013. Pursuing a relevant response in oral proficiency interview role plays. *Assessing second language pragmatics*, ed. by Steven J. Ross, and Gabriele Kasper, 288–310. Basingstoke: Palgrave Macmillan.

Ortega, Lourdes. 2011. SLA after the social turn: where cognitivism and its alternatives stand. *Alternative approaches to second language acquisition*, ed. by Dwight Atkinson, 167–80. London: Routledge.

Plough, India C., and P. S. Bogart. 2008. Perceptions of examiner behavior modulate power relations in oral performance testing. *Language Assessment Quarterly* 5. 195–217.

Reemann, Edith, Ene Alas, and Suliko Liiv. 2013. Interviewer behaviour during oral proficiency interviews: a gender perspective. *Eesti Rakenduslingvistika Uhingu Aastaraamat* 9. 209–26.

Roca-Varela, María Luisa, and Ignacio M. Palacios. 2013. How are spoken skills assessed in proficiency tests of general English as a foreign language? A preliminary survey. *International Journal of English Studies* 13. 53–68.

Roever, Carsten. 2011. Testing of second language pragmatics: past and future. *Language Testing* 28. 463–81.

Ross, Steven J. 2007. A comparative task-in-interaction analysis of OPI backsliding. *Journal of Pragmatics* 39. 2017–44.

Ross, Steven J., and Gabriele Kasper (eds). 2013. *Assessing second language pragmatics*. Bristol: Basingstoke.

Sacks, Harvey. 1992. *Lectures on conversation, volumes I and II*, ed. by G. Jefferson with introduction by E. A. Schegloff. Oxford: Blackwell.

Sacks, Harvey, Emanuel A Schegloff, and Gail Jefferson. 2014. A simplest systematics for the organization of turn-taking for conversation *Language* 50. 696–735.

Sato, Masatoshi, and Roy Lyster. 2012. Peer interaction and corrective feedback for accuracy and fluency development. *Studies in Second Language Acquisition* 34. 591–626.

Sandlund, Erica, and Pia Sundqvist. 2011. Managing Task-Related Trouble in L2 Oral Proficiency Tests: Contrasting Interaction Data and Rater Assessment. *Novitas-ROYAL – Research on Youth and Language* 5. 91–120.

——. 2013. Diverging task orientations in L2 oral proficiency tests - a conversation analytic approach to participant understandings of pre-set discussion tasks *Nordic Journal of Modern Language Methodology* 2. 1–21.

Schegloff, Emanuel A., Irene Koshik, Sally Jacoby, and David Olsher. 2002. Conversation analysis and applied linguistics. *Annual Review of Applied Linguistics* 22. 3–31.

Seedhouse, Paul. 2013. Oral proficiency interviews as varieties of interaction. *Assessing second language pragmatics*, ed. by Steven J. Ross, and Gabriele Kasper, 199–219. Basingstoke: Palgrave Macmillan.

Seedhouse, Paul, and Maria Egbert. 2006. The interactional organisation of the IELTS speaking test. Retrieved from IELTS Research Reports, http://www.ielts.org.pdf/Vol6_Report6.pdf.

Segalowitz, Norman, and Barbara F. Freed. 2004. Context, contact, and cognition in oral fluency acquisition: learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition* 26. 173–99.

Shohamy, Elena. 1990. Discourse analysis in language testing. *Annual Review of Applied Linguistics* 11. 115–31.

Sidnell, Jack, and Tanya Stivers. (eds). 2013. *The handbook of conversation analysis*. Chichester: Wiley-Blackwell.

Simpson, James. 2006. Differing expectations in the assessment of the speaking skills of ESOL learners. *Linguistics and Education* 17. 40–55.

Stansfield, Charles W., and Dorry Mann Kenyon. 1992. Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System* 20. 347–64.

Sullivan, JoAnn Hammadou. 2011. Taking charge: teacher candidates' preparation for the Oral Proficiency Interview. *Foreign Language Annals* 44. 241–57.

Tominaga, Waka. 2013. The development of extended turns and storytelling in the Japanese oral proficiency interview. *Assessing second language pragmatics*, ed. by Steven J. Ross, and Gabriele Kasper, 220–57. Bristol: Palgrave Macmillan.

van Compernolle, Rémi Adam. 2011. Responding to questions and L2 learner interactional competence during language proficiency interviews: a microanalytic study with pedagogical implications. *L2 interactional competence and development*, ed. by Joan Kelly Hall, John Hellerman, and Simona Pekarek Doehler, 117–44. Bristol: Multilingual Matters.

Van Lier, Leo. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly* 23. 489–508.

Walters, F. Scott. 2007. A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing* 24. 155–83.

——. 2009. A conversation analysis-informed test of L2 aural pragmatic comprehension. *TESOL Quarterly* 43. 29–54.

Wigglesworth, Gillian, and Cathie Elder. 2010. An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly* 7. 1–24.

Winke, Paula, and Susan Gass. 2013. The influence of second language experience and accent familiarity on oral proficiency rating: a qualitative investigation. *TESOL Quarterly* 47. 762–89.

Young, Richard, and Agnes Weiyun He (eds). 1998. *Talking and testing: discourse approaches to the assessment of oral proficiency (Studies in bilingualism 14).* Amsterdam: John Benjamins.

Appendix 1

Number of initial hits in the 15 examined journals in the period 2004–2014 (that is, all papers involving L2 oral proficiency testing).

Journal	**Hits (*n*)**
Annual Review of Applied Linguistics	2
Applied Linguistics	23
Classroom Discourse	1
Discourse Studies	0
ELT Journal	13
Foreign Language Annals	27
International Review of Applied Linguistics in Language Teaching	1
Journal of Pragmatics	6
Language Assessment Quarterly	8
Language Learning	2
Language Testing	58
Research on Language and Social Interaction	0
TESOL Quarterly	7
Text and Talk	0
Studies in Second Language Acquisition	15
Total	163