Language Testing 2009 26 (3) 325–339

Editorial

Are two heads better than one? Pair work in L2 assessment contexts

Lynda Taylor *University of Cambridge ESOL Examinations, UK*
Gillian Wigglesworth *University of Melbourne, Australia*

I Background

The move some years ago towards a more communicative approach in language teaching resulted in the increasing use of pair work in second language learning contexts, and this has been matched by the growth of paired assessments in testing contexts (see Hawkey, 2004). This in turn led to opportunities for the systematic collection and analysis of paired and group interactions and enabled the exploration of patterns of spoken interaction, taking into account variables such as age, gender, nationality, L1 of speakers, and so forth. The analysis of discourse, and in particular the detailed analyses provided through CA (conversation analysis), have proved particularly fruitful over the past 10–15 years, providing insights into talk-in-interaction and an enhanced understanding of the impact that working in pairs can have both on second language acquisition processes, and in the context of second language assessments.

The genesis of this special issue of *Language Testing* was the joint ILTA-AAAL symposium in 2007. A synergy and overlap between the interests and membership of the International Language Testing Association (ILTA) and those of the American Association of Applied Linguistics (AAAL) has existed for many years as a result of which the joint symposium was instituted some years ago by the two organizations. Designed as a meeting place for applied linguists and language assessment specialists, the symposium provides a shared discourse space in which to come together and discuss matters of mutual professional interest. The symposium now routinely takes place on the first or second day of the annual AAAL conference and is designed to generate ideas, cross disciplinary

DOI:10.1177/0265532209104665

boundaries, and disseminate research about issues and concerns in language policy, second language acquisition, language pedagogy and assessment, discourse analysis, and other disciplinary areas of applied linguistics.

For the 2007 AAAL conference, the joint symposium was entitled *Are two heads better than one? Pair work in L2 learning and assessment.* It featured papers that addressed this issue from both a second language learning perspective and from an assessment perspective. This special issue focuses, necessarily, on the assessment perspectives. It brings together a set of articles on the theme of pair work in the broad context of assessment, most of which are based upon papers delivered at the symposium. In assembling this collection of papers we seek to answer, at least in part, various questions:

- What does the research into pair work (both that which has been conducted in the past and the more recent research studies reported here) tell us about the benefits of pair work in assessment situations?
- Does it highlight any specific or significant disadvantages?
- Do the available research findings raise any implications for the language testing community in terms of our test theory and practice for the future?

II The potential benefits of pair work in the language learning context

There would appear to be clear advantages to pair work in both learning and assessment contexts. In the learning context, students are given more opportunities to actively use both their receptive and their productive language skills, including the opportunity to provide and obtain feedback from other students. In the assessment context, testing students in pairs (or groups), that is, with their peers, means that they have the opportunity to demonstrate their interactive skills in ways not generally available in more traditional testing formats, such as the one-on-one oral proficiency interview format, which typically involves a single test-taker with an examiner, who may combine their role as interlocutor-facilitator with that of assessor or rater.

From a research point of view, pair work in classroom teaching or in formal tests provides an ideal forum in which to gather and analyse the data of paired and group interactions so as to enable the exploration of varying patterns of interaction, with the potential to take into account variables such as age, gender, nationality, L1 of the speakers, and so on. The available research that has investigated pair

work presents a variety of different approaches to analysing paired discourse. Such analyses can provide us with valuable insights into, and thus an enhanced understanding of, the impact that working in pairs can have both on second language acquisition processes and in the context of second language assessments.

In relation to this we can reflect upon a variety of more specific questions. For example, what are the qualitative and quantitative features that distinguish paired test discourse from the discourse of the traditional one-on-one test format? Is there any evidence that working with a more proficient learner helps a less proficient learner? Can this be detrimental to the more proficient learner, or does being in the role of 'knower' or 'expert' actually enhance their own language performance? What impact does the more proficient learner have on the less proficient learner's performance? How do raters perceive the performance of the pair for the purposes of assigning scores, and how can pairs be assessed in ways that are equally fair to both participants? Other questions relate to whether and how the implications of working in pairs differ across spoken and written activities, and the contribution that different methodological approaches can make to our understanding of this phenomenon.

The empirical studies reported in this special issue explicitly set out to address a number of the questions listed above, adopting in the process a variety of methodologies. In addition, the papers stimulate wider questions relating to some fundamental issues for all who are involved in language testing and assessment. These issues include the following:

- the definition and operationalization of ability constructs;
- the development of assessment criteria and scales;
- approaches to rating and the training of raters;
- the interface between teaching, learning and assessment;
- and the limitations of methodologies and research outcomes.

We will reflect further upon some of these issues later in this introduction. First, however, we believe it will be helpful to briefly review some of the previous research literature relevant to pair work in language assessment, and the issues this raises, and then to offer a short summary overview of the five papers presented in this special issue.

III Research literature relevant to pair work in assessment

The assessment literature has long acknowledged that there are complex issues surrounding the interaction (usually spoken) that occurs in a performance-based proficiency test (Chalhoub-Deville, 2003;

Deville & Chalhoub-Deville, 2006; Luoma, 2004; McNamara, 1997; Swain, 2001; Weir, 2005). Whether the interaction involves a test taker and test examiner/rater in the traditional individual format, or a pair or group of test takers, the co-constructed nature of the interaction, and the fact that co-participants' contributions are inextricably linked, raises issues for language testers relating to construct definition, reliability and fairness. Various empirical studies can be cited showing how the behaviour of the test examiner/rater can affect test-taker performance (Brown, 2003, 2005; O'Sullivan, 2000; Ross & Berwick, 1992). The interlocutor effect may be due to personality (Berry, 2007; Bonk & Van Moere, 2004), proficiency level (Iwashita, 1996; Nakatsuhara, 2004, 2006), gender (Brown & McNamara, 2004; O'Sullivan, 2000) or acquaintanceship (or familiarity) (Katona, 1998; O'Sullivan, 2002). Other studies have confirmed that the interlocutor is not neutral in the interaction (He & Young, 1998; Lazaraton, 2002; Morton, Wigglesworth, & Williams 1997). Additional studies of test taker interaction with an examiner have highlighted the extent to which the resultant discourse structure can be significantly asymmetric (Lazaraton, 1996; Ross & Berwick, 1992; van Lier, 1989; Young & Milanovic, 1992) since it is the interlocutor who leads and controls the interaction. This imbalance in the power relationship between interlocutor and test taker is yet another factor that can shape the interaction that emerges during the testing event.

Growing awareness of some of the issues outlined above has led to the emergence and adoption of the paired or small group testing format (i.e. in which there is peer–peer interaction rather than or as well as examiner–test taker interaction). The paired or group approach is a potentially viable and effective alternative which addresses some of the concerns surrounding the more traditional approach. In addition, as the teaching and testing of speaking skills has been increasingly encouraged within language learning programmes, and as the practice has spread rapidly in classrooms around the world, the paired or group approach was perceived to offer a practical, time-efficient option for directly assessing large numbers of learners and providing them with feedback.

As take-up of the paired or small group approach has become more widespread, so the past decade has seen a growing body of literature emerge, both theoretically and empirically based, which reflects on this alternative format for testing second language ability and on its relative strengths and weaknesses (e.g. Berry, 1997, 2000; Bonk & Ockey, 2003; Bonk & Van Moere, 2002; Egyud & Glover, 2001; Galaczi, 2004, 2008; Iwashita, 1996; O'Sullivan, 2002;

Taylor, 2000, 2001; Van Moere, 2006). Much of this work appears to validate or support claims that the paired format can indeed resolve some of the weaknesses inherent in the traditional individual format. Some writers in the field, however, continue to highlight the issue of the joint construction of performance between test takers as problematic where the typical goal in testing is to assign each test taker an individual score (Foot, 1999; Fulcher, 2003; Norton, 2005). An extension of this issue is the challenge faced by raters when assessing spoken performance that is jointly constructed within a pair or group of test takers (see Fulcher, 2003; May, 2006).

In editing this special issue of *Language Testing* we hope to contribute to the emerging body of literature focusing on the use of pair work to test L2 proficiency.

IV Summary overview of the papers in this volume

The five papers in this special issue were all written by language testing researchers working with paired formats in a variety of contexts and for differing purposes. Some of the authors explore use of the paired assessment format as part of their postgraduate research training, and others as part of an experimental research agenda linked to a particular test or testing context. Some studies focus on the test takers, analysing their performance in the paired test format; others consider the rater (or raters) to explore issues associated with the rating process and product in the paired assessment context. Most of the papers are contextualized within the assessment of L2 speaking, but one paper is unusual in exploring the potential for using a paired format in L2 writing assessment. The studies reported here represent research conducted in a variety of contexts worldwide, including Australia, Canada and China, and though most focus on the testing of English as an L2, it is encouraging to be able to include one paper that explores the paired assessment of Spanish as a second language.

In the first paper, Lindsay Brooks examines and compares the interaction of adult ESOL test takers in two tests of oral proficiency, one where the candidates interacted with an examiner (individual format) and the other where they interacted with another student (paired format). Quantitative and qualitative analyses of the resulting score and discourse data suggested that the paired approach resulted in higher participant scores as well as a more complex interaction between the participants, including negotiation of meaning and

consideration of the interlocutor. Brooks frames her study within sociocultural theory (SCT) (Vygotsky, 1978; 1986) and in doing so introduces a thread which runs consistently throughout the papers, that is, the notion that 'the boundaries between social and individual functioning are quite permeable' (Wertsch, 1998, p. 110). This perspective chimes with McNamara's call for 'a renewed focus on the social dimension of interaction' (1997, p. 459) and echoes current thinking in the field about the fundamental socio-cognitive nature of assessment practices (Weir, 2005).

The second paper, by Larry Davis, reports on an experimental study conducted with a group of English learners at a Chinese university in which the focus was an examination of the influence of interlocutor proficiency on speaking performance in a testing context. The study was provoked by the oft-expressed concern that, from a measurement perspective, the paired oral risks being problematic because a partner's proficiency level may unfairly influence an examinee's performance or otherwise bias scores. After assessing students with partners of varying proficiency levels, the results showed that, on the whole, the majority of pairs produced collaborative interaction and that, given appropriate constraints, a difference in proficiency need not preclude the use of the paired format.

Lyn May's paper is the first of two papers to focus attention on the rater or raters, rather than the test takers, in a paired assessment. She identifies the need for improved definition and operationalization of interactional competence in paired or group speaking tests and reports on a study to explore the reactions of trained raters to pairs of test takers who oriented to asymmetric patterns of interaction in a discussion task. Drawing on the results of an analysis of candidate discourse and score data, complemented with a parallel analysis of rater recalls and discussions of the rating process, May concludes that raters perceive key features of the interaction during the test as mutual or shared achievements.

The second paper to focus on the rater (as opposed to the test taker) is by Ana Maria Ducasse and Annie Brown and reports the use of verbal protocol methodology with teacher-raters involved in evaluating the paired test discourse of a group of learners of Spanish as a Foreign Language in a university-based context. They take up the question of what raters actually focus on and attend to when rating paired interaction between test takers, and examine the extent to which the differences between interview and peer-peer talk are salient, describable and assessable from the perspective of the rater. As in May's paper, the focus here is on the process of defining the

construct of interaction and the necessity of doing this in order to develop appropriate assessment criteria and rating scale descriptors for judging performance outcomes.

The final paper in this special issue, by Gillian Wigglesworth and Neomy Storch, shifts our focus yet again, moving away from the role of pair or group work in assessing second language oral proficiency to its potential application in teaching contexts where formative assessments of second language writing ability are required. Their paper makes a welcome contribution to the relatively small body of research comparing the individual and collaborative production of written texts and the implications of this for approaches to assessment. Discourse analysis of the written product produced by each pair showed collaboration to impact on accuracy but not on fluency or complexity; a parallel analysis of the paired transcripts recorded during the collaborative writing process provides insights into how pairs work together and the foci of their endeavours. The paper ends by discussing the implications of the findings for classroom assessment practices.

V Some issues for consideration

The papers presented here help us to identify and problematize a number of fundamental issues in L2 proficiency assessment, some of a more theoretical nature and others touching upon more practical concerns.[1]

The issue of construct definition, that is, how the ability under scrutiny is to be conceptualized and described for the purposes of assessment, is a recurring theme throughout the papers. Several studies touch upon the nature of the L2 proficiency construct, specifically focusing on the need to better understand the nature of interactional competence so that it can be more adequately described and operationalized in assessment practices.

Approaches to defining a construct of L2 spoken language ability have evolved with construct definition becoming increasingly complex and comprehensive as it has sought to factor in communication-oriented as well purely linguistic skills. It is generally acknowledged that a direct, face-to-face speaking test format – as opposed to a semi-direct, tape- or computer-mediated approach – lends itself more

[1] We are grateful to one of the anonymous reviewers of this special issue for drawing our attention to some of the points discussed below.

readily to the elicitation of interactional competence (i.e. reciprocal interaction) for the purpose of evaluation. However, this approach also means that the interaction is likely to be unpredictable and subject to the influence of a range of variables that arise from its inherent reciprocity.

A traditional, measurement-oriented view in language testing typically considers variability in the language of the interlocutor in a face-to-face speaking test as an uncontrolled variable (resulting from age, gender, cultural background, L1 accent, L2 proficiency level, etc.), and thus an unwelcome threat to standardization, test reliability and fairness. For this reason some face-to-face speaking tests (especially those conducted formally on a large scale, such as the Cambridge ESOL speaking tests (see Taylor, 2003) or the Australian **access:** test (see Brindley & Wigglesworth, 1997)) make use of a formalized 'interlocutor frame'. This is a form of script which guides and constrains the examiner's talk (and timing) in order to offer all test takers a fair and equal opportunity to perform to their best. An interlocutor frame of this type may be designed to target and suit a particular proficiency level, requiring strict adherence on the part of the examiner; alternatively, it may be designed to be more flexible, so that it can be tailored by the examiner to the test taker's level in a test spanning the proficiency continuum.

An alternative view of interlocutor variability would be to regard it instead as part of the very ability construct that we are interested in measuring. According to this view, individual test taker characteristics such as age, gender, cultural background, L1 accent, L2 proficiency level, and so forth could be considered integral elements of communicative competence, and, as such, defined within the construct of interactional competence and operationalized within the paired speaking test. Several of the studies in this issue confirm the remarkable complexity and productivity of the interaction that can occur in the paired discourse of peer test takers. The findings also highlight the variable impact of differences at the level of the individuals involved, suggesting that interlocutor effects between peers are likely to be indirect and unpredictable rather than simple and consistent (cf. Brown and McNamara's 2004 findings on gender effect). Coping successfully with such interaction demands communicative skill and flexibility on the part of those involved. As one external reviewer commented in relation to Davis's paper on the effect of interlocutor proficiency: 'English as a lingua franca communication very frequently involves such demands on communicative skill.' Canagarajah (2006) makes a similar point when he presents an argument for multi-dialectal

testing in English. This is perhaps consistent with Deville and Chalhoub-Deville's (2006) view of variability as a potentially useful source of measurement information and validity evidence, rather than as a potential threat to reliability.

Several of the papers in this issue thus touch upon the need for our test theory and practice to account better for the joint construction of spoken language performance in a paired format test. The mediation which typically occurs between test takers in a paired group format can also be recognized as more fully representative of much language use in the world beyond the test, a world which Canagarajah suggests is characterized by 'postmodern globalization' (2006, p. 229) requiring an ever broader communicative repertoire on the part of language users. Adopting such a view may lead us towards richer construct definitions to underpin our test format and task design, and more sophisticated assessment criteria and rating scales to apply to the performance processes and products which result.

The final paper (on pair work in writing assessment) prompts us to consider construct definition and operationalization from a slightly different angle. This paper stimulates us to think about the extent to which similarities and differences can be discerned between paired interaction across spoken and written modes by considering the use of paired writing activities in formative assessment contexts in the classroom. There is an increasing sense among linguists and language educators that some of the traditional distinctions which have been recognized to hold between written and spoken production or performance are breaking down, in light of the blended forms that have emerged in modern electronic genres such as emails, weblogs, and various other forms of paired/group computer-mediated communication. Though still in written form, these genres display interactional features that share more in common with features of spoken than written interaction. In addition, language education has seen a move in recent years towards another sort of 'blending', one in which assessment activities are much more closely intertwined with teaching and learning activities, and in which different language skills are more closely integrated with one another. This point links to recent discussion on the *assessment for learning* (Rea-Dickins, 2008) and *dynamic assessment* (Lantolf & Poehner, 2008) paradigms. While a collaborative speaking task in a teaching and/or assessment context generates primarily oral output, a collaborative writing task may have the potential for generating an integrated sample of spoken and written output which could allow learners to produce texts that demonstrate their capacity for learning as well as providing a sample

of their written ability. There are interesting implications for the possible definition of a construct of interactional competence that is not constrained to either the spoken or written mode, but instead embraces both.

Discussion of issues of construct definition and operationalization leads naturally to the consideration of issues associated with assessment criteria and rating scales, and the nature of the actual rating process. Several of the papers highlight the question of how far existing criteria and scales developed for use in individual test formats are capable of capturing particular qualities of paired interaction formats, especially peer (or group) interaction. It may be that some criteria are more appropriately assessed for individuals in isolation, while other criteria are only meaningful if shared in common rather than owned individually. One implication of this would be that different parts of a test (e.g. monologue, collaborative task between peers) might merit different rating scales. Another is whether a task completion criterion is appropriate or not in a paired collaborative task. A further issue arises of whether individual or shared scores for interactional competence should be awarded to test takers in order to acknowledge the inherently co-constructed nature of paired interaction; and if so, how such an approach might be operationalized alongside, or instead of, the traditional practice of raters assigning each individual test taker their own individual score. A related issue is how many different assessment criteria and subscales it is reasonable and realistic to expect a rater to be able to handle in real-time during a speaking test event; and how are scores on multiple subscales best combined and reported to produce a meaningful outcome? These latter points highlight the practical challenges facing language testers when seeking to operationalize ever more complex and comprehensive definitions of the ability construct, particularly in large-scale assessment contexts.

A final area worth considering in this set of papers is what they can tell us about aspects of the research methodologies that are typically employed to investigate key research questions through an experimental paradigm, and in particular some of the methodological challenges and issues that researchers find themselves facing in this area.

The question has already been raised of how appropriate existing assessment criteria and rating scales may or may not be when applied to experimental test tasks or formats for which they were not originally designed. Researchers working in this area face a major challenge on this point. Do they 'borrow' pre-existing (and validated)

assessment criteria and rating scales from another context, perhaps applying them to performances from assessment tasks for which the criteria and scales were not designed and may not be well suited? Do they adapt pre-existing criteria and scales to better suit the experimental context, and if so, what does implications does this have for the effectiveness and integrity of the original scales? Do they design their own specific-purpose criteria and scales and if so, how far must these be trialled and validated prior to use in the main research? Scale design presents a particular problem in any research study that seeks to compare performance on individual tasks with that on paired tasks; the lack of an appropriate scale for one or other format may mean that essential features of performance cannot be credited and thus risks biasing the results in favour of one or other format.

A related issue concerns the selection and training of raters. If pre-existing criteria and scales are chosen for a study, then it follows that it is appropriate to adopt the established rater selection and training methods which accompany these. In practice, however, this is not always the case – indeed it may not be practically possible. However, if adequate account is not taken of the conditions which normally accompany use of the scales, then the reliability and validity of their application may be undermined as result. Absence of rater training – whether by design or by default – raises questions of reliability that need to be addressed. Deliberately avoiding rater training in order to achieve 'naïve rating' may promise fresh insights into the experience of assessing proficiency, but it may beg questions about how far a study's findings can generalize to a more formal assessment context. The challenge to the researcher is to determine what role rater selection and training should play in their study design and to justify this in appropriate ways.

Finally, studies in this special issue raise an interesting question over just how far their findings can be generalized beyond the immediate and relatively small-scale world of the experimental study to the wider world of language testing practice and practitioners. This raises complex questions of a broader epistemological and hermeneutical nature which cannot be fully discussed here. In theory, it is nevertheless generally accepted that knowledge and insights gained from research findings should have direct relevance and potential application to a wider practical context. In practice, however, this divide is not an easy one to cross and the extent to which experimental research findings can feed directly into practical, everyday language testing remains constrained. One reason is that practical

'real-world' testing, as opposed to language testing research, so often has to concern itself with so much more than issues of construct definition and operationalization, assessment criteria and rating scale development, rater selection and training, and so forth. Operational tests, especially large-scale commercial tests conducted on an 'industrial' scale, are usually located within a complex ecology comprising multiple, interacting factors, many of which are simply not present or relevant in language testing research studies – for example, sustainability issues to do with test production, delivery, processing; practical issues concerning test timing, security, cost, accessibility; organizational issues relating to personnel (e.g. developing and sustaining the rater cadre) or to management (e.g. the revision of an existing test, or development of its replacement). Such issues are mentioned here in passing since they are directly relevant to any discussion of generalization from the world of language testing research to the world of practical language testing. For example, how would it be possible in practice to reconcile awarding paired test takers shared scores for interactional competence with the need for them to receive the individualized scores typically required to achieve personal goals? What would this mean in both theoretical and practical terms? How far can language testing researchers offer real-world solutions?

VI Conclusion

As guest editors of this special issue of *Language Testing* we are most grateful to our fellow contributors for their willingness to spend time and energy drafting and revising their papers so that their studies (with their research questions, methodologies, results and conclusions) can be shared more widely for the benefit of others interested in the role and value of pair work in speaking and writing assessment. We regard the breadth and variety reflected in this set of five papers to be a strength and distinguishing feature of this special issue.

Finally, it is particularly gratifying to be able to bring to the attention of the wider language testing community research undertaken by those who are relatively new to the field and who will undoubtedly be the ones to take forward this important area of research and development in future years.

VII References

Berry, V. (1997). Ethical considerations when assessing oral proficiency in pairs. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment – Proceedings of LTRC 96* (pp. 107–123). Jyväskalä: University of Jyväskalä and University of Tampere.

Berry, V. (2000). An investigation into how individual differences in personality affect the complexity of language test tasks. Unpublished doctoral dissertation, King's College, University of London.

Berry, V. (2007). *Personality differences and oral test performance.* Frankfurt: Peter Lang.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20,* 89–110.

Bonk, W. J., & Van Moere, A. (2002, December). *L2 group oral testing: The influence of shyness/extrovertedness and the proficiency levels of other group members on individuals' ratings.* Paper presented at the AILA Congress, Singapore.

Bonk, W. J., & Van Moere, A. (2004, March). *L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores.* Paper presented at the Language Testing Research Colloquium, Temecula, California.

Brindley, G., & Wigglesworth, G. (1997) ***access:*** *Issues in language test design and delivery* Sydney: NCELTR.

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

Brown, A. (2005). *Interviewer variability in language proficiency interviews.* Frankfurt: Peter Lang.

Brown, A., & McNamara, T. (2004). "The devil is in the detail": Researching gender issues in language assessment. *TESOL Quarterly, 38,* 524–538.

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 3,* 229–242.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 9–25). Amsterdam: John Benjamins.

Egyud, G., & Glover, P. (2001). Oral testing in pairs – a secondary school perspective. *ELT Journal, 55*(1), 70–76.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*(1), 36–41.

Fulcher, G. (2003). *Testing second language speaking.* Harlow: Pearson.

Galaczi, E. (2004). Peer-peer interaction in a paired speaking test: The case of the First Certificate in English. Unpublished PhD dissertation, Teachers College, Columbia University, New York.

Galaczi, E. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89–119.

Hawkey, R. (2004). *A modular approach to testing English language skills: The development of the Certificates in English Language Skills (CELS) examinations*. Cambridge: UCLES/Cambridge University Press.

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). Amsterdam: John Benjamins.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51–65.

Katona, L. (1998). Meaning negotiation in the Hungarian oral proficiency examination of English. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 239–267). Philadelphia, PA: John Benjamins.

Lantolf, J. P., & Poehner, M. (2008). Dynamic assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education.* (2nd ed.). Volume 7: *Language testing and assessment* (pp. 273–284). New York: Springer Science+Business Media LLC.

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

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: UCLES/Cambridge University Press.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated recall. *Melbourne Papers in Language Testing, 11*(1), 29–51.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446–466.

Morton, J., Wigglesworth, G., & D. Williams. (1997). Approaches to validation: evaluating interviewer performance in oral interaction tests. In G. Brindley & G. Wigglesworth (Eds.), ***access:** Issues in language test design and delivery* (pp. 175–196). Sydney: NCELTR.

Nakatsuhara, F. (2004). An investigation into conversational styles in paired speaking tests. MA dissertation, University of Essex.

Nakatsuhara, F. (2006). The impact of proficiency level on conversational styles in paired speaking tests. *Cambridge ESOL Research Notes, 25*.

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal, 59*(4), 287–296.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*, 373–386.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277–295.

Rea-Dickins, P. (2008). Classroom-based language assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.).

Volume 7: *Language testing and assessment* (pp. 257–271). New York: Springer Science+Business Media LLC.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*(2), 159–176.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275–302.

Taylor, L. (2000). Investigating the paired speaking test format. *Cambridge ESOL Research Notes, 2*, 14–15.

Taylor, L. (2001). The paired speaking test format: Recent studies. *Cambridge ESOL Research Notes, 6*, 15–17.

Taylor, L. (2003). The Cambridge approach to speaking assessment. *Cambridge ESOL Research Notes, 13*, 2–4.

van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly, 23*, 489–508.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*, 411–440.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.

Weir, C. J. (2005). *Language Testing and Validation*. New York: Palgrave Macmillan.

Wertsch, J. V. (1998). *Mind as action*. Oxford: Oxford University Press.

Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 403–424.