

An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall

Lyn May

Nanyang Technological University

Abstract

This paper reports on the findings from an exploratory study in which rater orientations were examined through the use of stimulated verbal recall. Two trained and experienced raters of paired candidate discussion tasks produced retrospective verbal reports on six paired candidate discussion performances, producing a total of twelve retrospective verbal reports. Through an analysis of these verbal reports, it was possible to explore the raters' adherence to criterion and non-criterion features of the performances. It was found that trained and experienced raters attended to many non-criterion features of the paired candidate discussions, and as reflected in the focus of their comments, the raters differed in the extent to which they viewed the performance as co-constructed. These findings have implications for both the development of rating scales and the training of raters for paired candidate discussion tasks.

Address for correspondence: Lynette Anne May, National Institute of Education, 1 Nanyang Walk, Singapore 637616; Email: lynette.may@nie.edu.sg

Melbourne Papers in Language Testing, 1, 29-51.
ISSN 1327-0311 (print)

1. Introduction

Concerns over the validity of inferences which can be made from results of performances on high-stakes tests of oral proficiency continue to be voiced in the language testing literature (McNamara, 1996; Johnson, 2001; Brown, 2003).

In order to further explore the performance elicited by, and the rating of, tests of oral proficiency, research methodology which is more qualitative in orientation is being used with increasing frequency (Swain, 2001; Lazaraton, 2002; McNamara, Hill & May, 2002). Research tools including discourse analysis, conversation analysis and verbal protocol analysis now appear to be viewed not as mutually exclusive, but in some ways complementary to the previous reliance on a more positivistic approach to language testing research design.

These influences on language testing have created an intellectual climate which has encouraged researchers of performance speaking tests to explore the discourse elicited by particular tasks, candidate and interlocutor aspects, and the orientation and decision making of raters. The study reported in this article examines, through the use of stimulated verbal recall, the features of a paired candidate discussion task to which trained raters attend.

2. Background

Research into performance tests of speaking: focus on paired candidate interactions

Following a research agenda owing much to both van Lier's (1989) provocative analysis of a traditional language testing "conversational" interview and McNamara's (1997) questions regarding the co-construction of performance in interactive speaking tests, language testing researchers have closely examined aspects of two high-stakes tests of oral proficiency: the Oral Proficiency Interview (OPI) and the International English Language Testing System (IELTS) interview. In contrast, the paired candidate interaction, which is also used in high-stakes tests, has received far less attention in the language testing literature.

One of the earlier published studies of paired candidate interactions was that of Iwashita (1998), who compared the impact on candidates' scores and discourse when paired with an interlocutor of a similar and different proficiency level. The participants were twenty adult learners of Japanese. She found that although the proficiency of the interlocutor did impact on the quantity of discourse elicited through the task, it did not seem to significantly change scores given to candidates. In addition, test-taker feedback indicated that "candidates prefer the NNS-NNS interaction mode to the NS-NNS mode as they find it less threatening" (p.52). Candidate preference for the paired candidate interaction was also reported by Egyud and Glover (2001), Taylor (2001) and May (2000).

Ikeda (1998) explored the paired candidate interaction from a sociocultural perspective. Through a study of five "paired learner interviews", involving teenage Japanese students of English, he found that this testing task offered the candidates opportunities not only to negotiate meaning, but also to "take initiative to learn new knowledge and incorporate it into their respective private worlds" (p.71). Ikeda allowed candidates to select their interlocutor, and cautioned against the "risk of pairing linguistically compatible learners who may be incompatible personality-wise" (p.93).

Paired tests of spoken language were heavily criticized by Foot (1999), who was particularly concerned about the lack of published research on this task, which had already been incorporated into several UCLES high-stakes speaking tests. A potential problem highlighted by Foot (1999) was the prospect of candidates of differing spoken proficiency levels being disadvantaged in a paired candidate speaking test: "unless the candidates are well-matched, their attempts to sustain a discussion are likely to be, and often are, faltering and desultory, and the outcome for them a sense of frustration rather than achievement" (p.40). Foot (1999) was also concerned with the prospect of mutual incomprehensibility if both candidates had pronunciation problems, or even particularly strong accents. In comparing the paired candidate tests to the traditional interview, Foot (1999: 40) concluded "it is difficult to see how a discussion between two inexpert users, struggling to overcome their own limitations, and attempting to decipher the opacities of the other, is compatible with providing candidates with the optimum conditions for showing what they can do". Referring to the perception that

candidates preferred the paired speaking test and felt more relaxed while performing this task, Foot cautioned that candidate preference is not a sufficient reason to incorporate a testing task into a high-stakes test.

Responding to Foot's criticism on behalf of UCLES, Taylor (2001) reported the results from two internal studies which had been undertaken in order to compare paired and one-to-one speaking test formats. The paired speaking test format was shown to elicit more *informational functions* and *managing interaction functions* than the one-to-one interview. In addition, whereas *informational functions* made up approximately 80% of the candidates' discourse in the one-to-one interview, they accounted for only 55% of the candidates' discourse in the paired speaking test format (p.16). From this Taylor concluded that paired speaking tests have the potential to be more symmetrical and genuinely interactive than the traditional one-to-one interview.

Concern over the lack of a focused research agenda into pair and group speaking tests was also expressed by Swain (2001): "given that small group testing occurs in even one high-stakes test, as well as its reasoned use, it is surprising that so little validation work has been carried out" (p.277). Linking McNamara's (1997) concerns relating to the co-constructed discourse being rated as the product of the individual, she recommended that candidate discourse be examined, so that a deeper understanding could be reached about exactly what was being elicited through pair and group test tasks, which could "provide test-developers with targets for measurement" (p.297).

Following the publication of Swain's (2001) article, a number of studies on paired candidate speaking tests appeared. The research that most directly addressed Swain's call for closer examination of discourse was that of Galaczi (2003; 2004), Brooks (2003; 2004) and Lu (2003a; 2003b). The impact of the pairing of candidates was further explored by O'Sullivan (2002), Nakatsuhara (2004) and Norton (2005).

Swain (in Fox, 2005) strongly advocates collaborative testing tasks, and argues that "if students were taught how to scaffold and what negotiation of meaning really is, then those characteristics could be looked at in a testing situation, and I think we would be, for most learners, biasing for best" (p.242).

Rater orientation and decision making in tests of speaking

Pollitt and Murray's study (1996) is of particular relevance to the rating of paired candidate interactions. While the raters viewed two individual performances involving different candidates, they were then asked to decide which performance was *better*. Following this, they instructed raters to explicitly compare and contrast the performances of the candidates. Through this approach Pollitt and Murray concluded that while some raters followed a *synthetic* process of rating which appeared to be more intuitive, other raters were engaged in a more analytical approach. Another finding was that where candidates in a pair were of differing proficiency levels, "the criteria judges focused on were generally those associated with the lower-level candidate of the pair" (p.86). If this finding is replicated on a larger scale, it has serious implications for the use of paired candidate interactions in high-stakes tests.

The direction of three recent studies into rater orientation and decision making in tests of speaking reflects the extent to which language testing is incorporating research tools from other traditions, as one of the notable features that these studies have in common is the use of protocol analysis.

Brown (2000) used stimulated verbal recall (DiPardo, 1994) in a study of rater decision making within the context of the IELTS interview. She found evidence to suggest that the decision making of the eight trained raters in this study could also be categorized using Pollitt and Murray's (1996) synthetic and analytic approaches. She concludes that rating "is and will always remain an 'imprecise science' and raters deserve to be given credit for their attempts to make sense of the interaction and quantify as they are required to do" (p.81).

In a study involving the First Certificate in English (FCE) speaking test, Orr (2002) used retrospective verbal reports from 32 trained raters to help interpret test scores. He found that "raters did not heed the same aspects of the assessment criteria, and heeded a wide range of non-criterion relevant information" (p.143). As a result, raters could come to the same rating, but for different reasons. Orr (2002) reaches harsher conclusions than Brown (2000) about the trained raters who participated in his study: "the verbal reports of many raters show difficulty in adhering to the assessment criteria. There is

also evidence that raters do not understand the model of communicative language ability on which the rating scales are based" (p.153). He calls for more focused rater training, and further examination into the rating scales, in order to ascertain whether adjustments need to be made.

Brown, Iwashita and McNamara (2005) used both retrospective verbal reports from raters and a discourse analysis of spoken language elicited through tasks designed for the Test of English as a Foreign Language Test of Spoken English (TOEFL TSE) in order to explore the extent to which raters' perceptions matched the actual discourse features of a performance. In keeping with findings from previous studies, they suggest that individual raters will find different features of a performance more salient, perhaps reflecting their individual frames of reference and experience.

Issues inherent in the use of verbal reports in language testing research

Verbal reports are seen as an avenue through which to gain insights into an individual's thought and decision making processes which it is not possible to obtain through other research tools (Lumley, 2000). Yet Lumley (2000: 305), who used think aloud protocols in a study of rater decision making in the context of rating academic writing, also cautions against accepting the results from think aloud protocols at face value, pointing out that "a lack of mention of a particular feature or features by a rater is no indication that the feature was not observed and noted. Raters explicitly make the point that far more passes through their minds than they can ever articulate".

Traditional areas of concern relating to the validity of verbal protocols focus on the temporal relationship between the verbal report and the action or event which is used to elicit it. Cohen (1987) cautions that people can forget salient aspects of their thought processes almost as soon as a thought has passed through their mind: "It appears that the bulk of the forgetting occurs right after the mental event. Thus, data from immediate retrospection may only be somewhat more complete than data from delayed retrospection (p.84)."

It is also recommended that informants be trained prior to the production of verbal reports: the failure to do so is seen as a threat to validity (Faerch & Kasper, 1987). Yet Ericsson and Simon (1987) caution against the uncritical acceptance of verbal reports, even with trained informants, as “immediate and direct observations of those [cognitive] processes, veridical and uncontroversial” (p.24).

Smagorinsky (2001) also expressed concern about the validity of verbal protocol analysis as a research tool. From a cultural perspective, he maintains that “if thinking becomes rearticulated through the process of speech, then the protocol is not simply representative of meaning. It is, rather, an agent in the production of meaning” (p.240). Other concerns of Smagorinsky (2001) are the social nature of the verbal protocol and its “hidden dialogicality” (p.238) which he believes are not acknowledged by cognitive psychologists from whom the methodology originated.

Implications for research

From the review of recent research on rater decision making in tests of oral proficiency, it is clear that despite some concerns related to validity, verbal reports have given valuable insights into rater orientations. Thus I decided to use stimulated verbal recall as the research tool with which to explore rater orientations on a paired candidate discussion task.

The research question that I will report on in this article is: which features of the performance do raters attend to when assessing a paired candidate interaction?

3. Methodology

Raters

Two experienced raters from a university language centre in Australia volunteered to take part in the research. These raters were trained to rate paired candidate discussions, as this task is used in a high-stakes English for Academic Purposes test in the centre. In order to keep the raters’ personal details confidential, the male assessor will

be referred to as Rater 1, and the female assessor as Rater 2 in this article. Table 1 presents the shared characteristics of the raters.

Table 1 Shared characteristics of the raters

Shared characteristics of the raters
<ul style="list-style-type: none"> • speakers of English as a first language • post-graduate Teaching English to Speakers of Other Languages (TESOL) qualifications • five or more years of English for Academic Purposes (EAP) teaching experience • recent teaching experience with adults • two or more years experience in rating paired candidate interactions

Candidates

The twelve candidates (six male and six female) were all scholarship holders from China. They ranged in oral proficiency level from intermediate to advanced, and in age from 18 to 20. They volunteered to participate in the paired candidate speaking tests following the conclusion of a six month intensive EAP course given at a tertiary institution in Singapore.

Paired candidate discussion task and performances

The speaking task was a structured discussion task, with a problem/issue presented and three possible solutions for discussion. The candidates had up to five minutes of planning time to prepare for the discussion. In the actual high-stakes test situation, candidates are given a theme-based reading test, followed by a lecture which continues the theme, and are then given one hour to write an essay also based on the theme. The final task is the paired candidate interaction, where the issue discussed is also related to the theme of the test.

As it was not possible to replicate the entire theme-based test, several days prior to the task, candidates in the study were given two readings related to each of the issues for discussion, which were human cloning (Task A) and genetically modified food (Task B). They were instructed to read the texts before their allocated test time, and

to bring the readings to the testing venue. Each candidate took two forms of the paired speaking test: one with a partner at a similar level, and one with a partner of a different level. Parallel tasks (Task A and Task B) were used. Performances were both video- and audio-taped.

Rating scales

An analytic rating scale had been devised for the rating of the paired candidate interaction by the group of test developers responsible for implementing the paired candidate interaction task at the university language centre in Australia. This scale consisted of five categories: *Fluency*; *Accuracy*; *Range*; *Effectiveness*; and *Overall*. There are descriptors for band levels 1-5 within each of these categories.

Generating the stimulated verbal recalls

Raters were instructed to trial one verbal protocol, in order to experience producing a verbal report. They then produced retrospective verbal protocols on a set of six selected paired candidate performances. After viewing each performance together the first time, raters gave their rating. They were then instructed to view the performance individually, stopping the video at any point that they felt something was said or happened that was important /noticeable/helpful to influence their rating, and comment on it. These verbal protocols were audio-taped.

The reason for selecting six of the twelve interactions for verbal protocol reports was that it was important to limit the amount of data being generated in an exploratory study. The basis for the selection was that these performances constituted a representative sample, in that they included a range of pairings with respect to gender and oral proficiency levels.

Data transcription, segmentation, coding and analysis

The twelve paired candidate interactions and twelve individual verbal protocol reports were transcribed using orthographic transcription conventions from Atkinson and Heritage (1984, in Lazaraton, 2002). These include the use of brackets ([]) for overlapping talk; a colon (:) for a lengthened sound or syllable; timed

pauses and capital letters (CAPS) for a word or sound that is emphasized.

Segmenting the verbal protocols

The verbal protocols were segmented according to the imperative that “each segment should be representative of a single, specific process” (Green, 1998). This means that one review turn could generate several segments. An example from the data is:

1-01-26-B

immediately noticing that Jun really has quite intractable pronunciation problems / B, AC:PRO -ve

even for an experienced teacher like myself I would really have to struggle to work out what he is saying / B, RR:TEA -ve

1-01-26-B indicates this review turn is from Rater 1’s verbal protocol on paired candidate interaction 1, and that the rater stopped the tape at a point which corresponds to Line 26 in the transcribed paired candidate interaction, and is referring to Candidate B. The one review turn generates two segments, each of which are coded. B, AC:PRO -ve indicates that the remark is about candidate B (sometimes both candidates are referred to in the same turn), and is concerned with Accuracy in pronunciation, and is negative. B, RR:TEA -ve indicates that the remark is about Candidate B, and that this segment is coded as a Rater Reflection, where the rater relies on his experience as a teacher, and is negative.

If one review turn contained several segments, two of which although separated, clearly referred to the same aspect of performance, the second segment would not be counted as separate. This decision impacts on the frequency count, but is in accordance with Brown *et al* (2005: 14), who redefined an Ideas Unit to incorporate non-continuous speech as “a single or several utterances, either continuous or separated by other talk but falling within the same turn, with a single aspect of the performance as its focus”.

An example from the present study is:

1-01-56-A

OK I hear Shen there continuing to make a reasonable attempt at expressing complex ideas / A, TR:COM+ve

a few incidental grammatical mistakes "the science" struggling with the present perfect / A, AC:GRA-ve

running up against limitations of vocabulary because once again the only word he can come up with is "trouble" / A, RA:VOC-ve

but nevertheless there's a fairly impressive movement there towards being able to express complex ideas / *Not coded separately, as this is a continuation of first comment on complexity*

While most segments were easily distinguishable, some were problematic. Hence, it is important to have a co-segmenter to check inter-segmenter reliability, in addition to inter-coder reliability, in order to have a less idiosyncratic representation of the data.

Coding the verbal protocols

Devising the coding key

As Green (1998: 68) states, the absence of agreement as to what exactly constitutes the "precise nature of the coding categories that may be used for the analysis of verbal report data" is problematic, and the consequence of this is that "two researchers may independently develop different schemes for the analysis of the same body of data". Although Green (1998) does not feel that this invalidates the technique, she cautions that this variability will inevitably affect the inferences that can be drawn from the results. To prevent possible invalidity of this technique, Green (1998) suggests that a balance must be maintained between the researcher's desire for coding that reflects every nuance of a verbal report, and the need to establish inter-rater reliability. If coding categories are too broad, inter-coder reliability may be higher, but it would be more difficult to make meaningful inferences from the data. Although more specific coding categories may yield insights, they will probably result in lower inter-coder reliability.

With this in mind, I read through the verbal protocols several times before beginning to note categories of information that raters commented on. In addition to these, I used the criteria from the rating scales, which are referred to as criterion aspects of the rating. From this I compiled a set of criterion and non-criterion codes to represent rater's comments, which are presented in Tables 2, 3, 4 and 5.

From my first attempt to code, it became clear that the categories of evaluative response of the rater (+ve = positive response to candidate's performance; -ve = negative response to candidate's performance; N = neutral response/ non-evaluative response) were insufficient to represent the nature of the raters' comments on candidate performance in the paired candidate interaction. Because two candidates were involved, raters made many inter-candidate comparisons, which I felt were important to identify, as they reflected the orientation of the rater. Raters also made intra-candidate comparisons, ("her pronunciation is better now than at the beginning"), which I felt needed a separate coding symbol.

Thus I added another four symbols to the coding scheme in order to distinguish comments that were not strictly about features of an individual's performance at a specific point in time:

S = inter-candidate comparison, finding similarities; D = inter-candidate comparison, finding differences; C = intra-candidate performance, comparing an aspect of one candidate's performance over time; P = a comparison of an aspect of both candidates' performance over time.

Inter-coder reliability

A colleague experienced in coding data from verbal reports independently coded four rater protocols (33% of the total). After she had completed the coding, I compared our results, and we had an agreement of 83% in regards to the allocated codes. We then had a meeting to discuss the segments that we had coded differently. After agreeing on which code would be adopted, and in certain cases further segmenting the review turns, I then recoded all twelve individual rater protocols in light of our discussion.

Frequency counts

After the final coding, I tallied all 416 segments from the verbal protocols according to their codes. This enabled me to see rater tendencies more clearly, particularly in broad terms including the number and proportion of rater comments per category, the number and proportion of negative, positive and neutral comments made by raters, and the extent to which raters made inter- and intra-candidate comparisons.

4. Results and discussion

When coding the retrospective verbal protocol reports, it immediately became apparent that raters were noticing many features of the performance which were not mentioned in the rating descriptors. Rater comments were coded and then tallied, allowing for an analysis of rater orientations.

Attention to criterion features of the performance

Table 2 presents the criterion aspects of the performance commented on by the raters which were coded. Different categories of the rating scales appeared to be more salient to each rater. Whereas Rater 1 made many comments on aspects of accuracy, particularly relating to grammar and pronunciation, Rater 2 appeared to pay more attention to aspects of fluency, particularly hesitation, and also vocabulary range. This may reflect the individual frames of reference which each rater brings to the rating experience, incorporating their beliefs about language proficiency and possibly their orientations as language teachers. The grouping together of grammar, pronunciation and vocabulary under the category of "Accuracy", in addition to grammar and vocabulary grouped together under "Range", could also have led to the divergent focus of the raters.

Table 2 Criterion aspects of the performance

Criterion aspects	Details
FLUENCY	Fluency, mentioned in general Speed of delivery Hesitation Repetition
ACCURACY	Accuracy mentioned in general Accuracy - pronunciation Accuracy - vocabulary Accuracy - grammar Accuracy - self-correction
RANGE	Range mentioned in general Range - vocabulary Range - grammar
EFFECTIVENESS	Effectiveness mentioned in general Understands interlocutor's message Able to respond to interlocutor Uses communicative strategies
OVERALL	Use of descriptive, explanatory, evaluative and speculative language.

Aspects of the performance which raters regarded as criterion

It was interesting to note that raters appeared to have "fleshed out" the criteria in the band descriptors with features that were not explicitly mentioned in the band descriptors, but which from the content and context of their comments, raters clearly regarded as salient to the categories in the rating scales. Rater 2 noted the use of idiomatic and "natural" language favourably, whereas Rater 1 did not comment on this feature in the verbal protocols. Although both raters mentioned intelligibility to the rater (as opposed to the interlocutor, which is explicitly stated in the criteria), Rater 2 commented on this feature more than Rater 1. Both raters noticed moves to control the paired interaction, which were generally viewed negatively as manifestations of dominance, and moves to manage the discourse, which they viewed positively. Raters also commented on a candidate "helping out" another. Although this was always

mentioned in a positive way with respect to the candidate who was “helping”, I wonder to what extent raters incorporated into their rating the fact that help was needed in the first place. Table 3 presents aspects of the performance which the raters regarded as criteria which were coded.

Table 3 Incorporation of features not explicitly mentioned in band descriptors into the criteria

Features	Details
RANGE	Use of idiomatic language, and that described as “natural”
EFFECTIVENESS	Able to paraphrase own and partner’s ideas Able to express own ideas Intelligibility to rater Controls/ manages interaction Helps partner out

It is clear from the above coding of comments that raters have incorporated additional features which they feel are salient to the rating criteria. This could indicate the need for the rating scales to be revised and/or more comprehensive rater training, as each rater appears to interpret the criteria in a different manner.

Attention to non-criterion features of the performance

More than 30% of rater comments alluded to non-criterion aspects of the performance. Two main categories of non-criterion features of the performance noticed by raters emerged. These were broadly coded as Rater Reflection, and Task Realisation. The finer codings within these two categories are shown in Tables 4 and 5. Table 4 presents the comments made by raters which were categorised as rater reflection which were coded.

Table 4 Rater reflection

Rater reflection
Rater reflection in general
Reflecting on whether an inaccuracy impacts on the rating
Reflection on rating behaviour
Reflecting on own status/experience as a teacher in relation to the performance
Reflecting on candidate's proficiency level in general
Reflecting on candidate's language development
Reflecting on whether candidate is performing to his/her potential
Reference to "native speaker" level of competence
First impression of candidates
Matching of candidates
Confidence level of candidate
Sense of humour of candidate
Extent to which candidates are "warming" to the task
Ability (other than language) of the candidate
Voice quality of the candidate

Although both raters made reference to non-criterion features coded as rater reflections, differing patterns emerged. Rater 1 speculated about the ability of the candidates in areas other than language, and noted whether candidates were "well matched". Rater 2 frequently commented on an aspect of the performance which she termed "warming up" to the task, but Rater 1 did not mention this. Both raters reflected on the extent to which inaccuracies in grammar and pronunciation influenced their rating decisions. Table 5 presents the comments made by raters which were categorised as pertaining to task realization which were coded.

Table 5 Task realisation

Task realisation
Understanding the issues
Completing the task
Organisation of ideas, use of discourse markers
Extended discourse from a candidate
Quality of ideas
Complexity of ideas
Relevance of ideas
Logic of ideas
Analysis of arguments
Substantiating ideas (with examples etc.)
Ability to summarise the discussion
Reference to the readings
Extent to which the interaction resembles authentic discussion

Although the paired candidate interaction is an “integrated” task, which is assumed to be cognitively more demanding (Brown *et al*, 2005: 1) the rating descriptors do not explicitly address the content of the discussion: it is only seen as a vehicle through which to elicit a sample of speech in order to make judgements about a candidate’s linguistic ability, as if this were somehow a separate entity from the ideas themselves. Raters made comments about the candidates’ understanding of the issues involved, their task completion orientation, the quality of the ideas presented, incorporation of information from the readings into the discourse, and the extent to which the interaction resembled a “real” discussion.

It is interesting that the raters commented on features more commonly associated with academic writing than speaking. Rater 2 commented generally on the quality of ideas, whereas Rater 1 quite systematically referred to the complexity, relevance and logic of candidates’ ideas, in addition to the ability to analyse arguments. This could reflect a different understanding on the part of the raters of what the task was designed to elicit.

A clear difference emerged in rater orientation with regard to the integration of information from the readings into the discourse: Rater

1 referred to it negatively, as he felt it detracted from fluency; whereas Rater 2 commented positively on candidates' use of ideas from the readings, regarding this as evidence of the ability to synthesize information from various sources.

Overall rater orientations

Clear tendencies in rater orientation emerged through analysis of the verbal protocols. While 67% of Rater 1's comments were positive, Rater 2 had an almost equal proportion of positive and negative comments. This could reflect a tendency of Rater 1 to be a more "lenient" rater, or it could be the result of different interpretations of the rating scale.

The extent to which the raters differed in their acknowledgement of the co-construction of the performance is an area of interest. Rater 1 tended to comment on candidates individually, with only 13% of his comments involving inter-candidate comparisons, whereas 34% of Rater 2's comments were inter-candidate comparisons. This might be due to Rater 1 being more experienced in using the rating scale, and thus more focussed on the scale, rather than inter-candidate comparisons, but might also be a reflection of Rater 2's perception of the discourse as co-constructed, which would cause difficulty in conceptualising and rating the performance as if it were the manifestation of two distinct "solo" performances.

Rater 1 made constant comments as the interaction progressed, whereas Rater 2 was more likely to make overall summary comments about candidates at the end of an interaction. This could reflect different decision making processes: Rater 1 was more analytical; Rater 2 was more impressionistic and holistic.

5. Conclusion

The small scale of this exploratory study lends itself more to the generation of possibilities than conclusions. The adherence of trained and experienced raters to non-criterion aspects of the performances is a concern, and may indicate the need for further rater training and/or the revision of the rating scales used to rate the paired candidate interaction.

Further research is needed to establish which features of a complex, integrated performance test of speaking raters actually heed, and how they reach their rating decisions. There is also a need to acknowledge the difficulty of the task facing raters when attempting to reconcile aspects of complex paired candidate interactions with rating scales and their own frames of reference as both teachers and raters.

Of particular interest is the extent to which a rater acknowledges the co-construction of the paired candidate performance, and the impact this has on the final rating. The rating scales require raters to view the paired speaking test as if it were the product of two solo, quite distinct performances, which ignores the inherent co-construction of the performance. If one candidate's performance is adversely affected by, or compared with, his/her partner's, and the two are of different levels, issues of ethical testing could arise, particularly with respect to high-stakes tests.

The rating scales also lacked descriptors relating to the quality and quantity of ideas that were being expressed, focussing only on the linguistic aspect of the task. This is questionable when dealing with academic speaking tasks.

With regard to research methodology, the inherent subjectivity of the analysis of verbal protocols, in terms of deciding on idea units and the coding of the protocols is an area of concern. The application of cultural theory to verbal protocol analysis is another area that requires further exploration. If, as Smagorinsky (2001) asserts, there is a hidden dialogicality inherent in the production of a verbal report, it is possible that the protocoller is, at least subconsciously, addressing the researcher, rather than producing a report that reflects the response of the rater to the original performance. It is thus possible that the protocoller may be tailoring his/her comments to meet the perceived expectations of the researcher. Interviewing the protocollers after the production of the verbal reports may yield insights into whether they are, at some level, engaged in a dialogue with a researcher. Interviews can also be aimed at further exploring the extent to which protocollers are able to report on all the aspects that they heed in a performance.

The use of verbal protocol analysis is becoming increasingly common in studies of oral proficiency testing. While verbal protocols have the

potential to generate rich data which can offer insights into rater orientations, researchers who utilize this methodology should be aware of the areas of concern associated with it.

References

- Brooks, L. (2004). Insights into the construct(s): Paired oral proficiency testing. Handout on paper presented at the 26th Language Testing Research Colloquium, Temecula, United States.
- Brooks, L. (2003). An investigation of the interactions in paired oral proficiency testing. Handout on paper presented at the 25th Language Testing Research Colloquium, Reading, United Kingdom.
- Brown, A. (2000). An investigation of rater's orientation in awarding scores in the IELTS interview. In R. Tulloch (Ed.) *IELTS Research Reports*, 3, (pp. 1-19).
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20, 1-25.
- Brown, A., Iwashita, N., & McNamara, T. (2005). *An Examination of Rater Orientations and Test-taker Performance on English-for-Academic-Purposes Speaking Tasks*. TOEFL-MS-29. Educational Testing Service: Princeton, NJ.
- Cohen, A.D. (1987). Using verbal reports in research on language learning. In C. Faerch & G. Kasper (Eds.), *Introspection in Second Language Research*. (pp. 82-95). Clevedon: Multilingual Matters.
- DiPardo, A. (1994). Stimulated recall in research on writing: an antidote to "I don't know, it was fine". In P. Smagorinsky (Ed.), *Speaking About Writing: Reflections on Research Methodology*. (pp. 163-184). Thousand Oaks, California: Sage.
- Egyud, G. & Glover, P. (2001). Oral Testing in pairs: A secondary school perspective. *ELT Journal*, 55, 70-76.
- Ericsson, K.A. & Simon, H.A. (1987). Verbal reports on thinking. In C. Faerch & G. Kasper (Eds.), *Introspection in Second Language Research*. (pp.24-53). Clevedon: Multilingual Matters.

- Faerch, C. & Kasper, G. (1987). From product to process - introspective methods in second language research. In C. Faerch & G. Kasper (Eds.), *Introspection in Second Language Research*. (pp. 5-23) Clevedon: Multilingual Matters.
- Foot, M.C. (1999). Relaxing in pairs. *ELT Journal*, 53, 36-41.
- Fox, J. (2005). Biasing for the best in language testing and learning: An interview with Merrill Swain. *Language Assessment Quarterly*, 1, 235-251.
- Galaczi, E.D. (2003). Interaction in a paired speaking test: the case of the First Certificate in English. *UCLES Research Notes*, 14, 19-23. http://www.cambridge-efl.org/rs_notes/index.cfm
- Galaczi, E.D. (2004). Peer-peer interaction in a paired speaking test: the case of the First Certificate in English. PhD dissertation: Columbia University.
- Green, A. (1998). *Verbal protocol analysis in language testing research. Studies in Language Testing* 5. Cambridge: Cambridge University Press/UCLES.
- Ikeda, K. (1998). The paired learner interview: A preliminary investigation applying Vygotskian insights. *Language, Culture and Curriculum*, 11, 71-96.
- Iwashita, N. (1998). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing*, 5, 51-65.
- Johnson, M. (2001). The art of non-conversation: A re-examination of the validity of the Oral Proficiency Interview. New Haven: Yale University Press.
- Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests. Studies in Language Testing* 14. Cambridge: Cambridge University Press/UCLES.
- Lu, Y. (2003a). Insights into the FCE Speaking Test. *UCLES Research Notes*, 11, 15-19. http://www.cambridge-efl.org/rs_notes/index.cfm
- Lu, Y. (2003b). Test-takers' first languages and their discursive performance in paired-format OPT. Handout on paper presented at the 25th Language Testing Research Colloquium, Reading, United Kingdom.

- Lumley, T. (2000). *The Process of the Assessment of Writing Performance: The Rater's Perspective*. PhD thesis. The University of Melbourne.
- May, L. (2000). Assessment of oral proficiency in EAP programs: A case for pair interaction. *Language and Communication Review*, 9, 13-19.
- McNamara, T. (1996). *Measuring second language performance*. Harlow: Addison Wesley Longman.
- McNamara, T. (1997). 'Interaction' in second language performance assessment: whose performance? *Applied Linguistics*, 18, 444-446.
- McNamara, T., Hill, K. & May, L. (2002). Discourse and Assessment. *Annual Review of Applied Linguistics*, 22, 221-242.
- Nakatsuhara, F. (2004). An Investigation into Conversational Styles in Paired Speaking Tests. MA dissertation: University of Essex.
- Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT Journal*, 59, 287-297.
- O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19, 277-295.
- Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System*, 30, 143-154.
- Pollit, A. & Murray, N.L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*. (pp.74-91). Cambridge: Cambridge University Press.
- Smagorinsky, P. (2001). Rethinking protocol analysis from a cultural perspective. *Annual Review of Applied Linguistics*, 21, 233- 245.
- Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, 18, 275-302.
- Taylor, L. (2001). The paired speaking test format: recent studies. *UCLES Research Notes* 6, 15- 17. http://www.cambridge-efl.org/rs_notes/index.cfm.

van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral Proficiency Interviews as conversations. *TESOL Quarterly*, 23, 480-508.