Crosthwaite, P., Boynton, S. & Cole, S. (2017). Exploring rater conceptions of academic stance and engagement during group tutorial discussion assessment. *English for Academic Purposes*, to appear.

Exploring rater conceptions of academic stance and engagement during group tutorial discussion assessment

Abstract

The present study uses concurrent think-aloud verbal protocols alongside post-hoc interviews to explore how six teacher-raters determine a students' ability to explain academic concepts and argue for an academic stance supported by sources during a 25-minute group tertiary academic tutorial oral assessment. We explored how the raters arrived at decisions regarding the quality of students' academic stance and engagement in light of difficulties with rater attention in real-time, L2 language concerns, assessing engagement in a group oral setting, and the use of spoken citation to support speakers' claims. Substantial differences in rater practice, beliefs and interpretation of assessment criteria were all found during the assessment of student performance, confirming a number of difficulties faced by raters assessing the academic ability of multiple participants over lengthy extended, interactional discourse. The findings shed real-time conceptions of (un)successful academic stance and engagement in group oral contexts, as well as confirm the usefulness of verbal protocols in revealing previously hidden complications for group oral assessments in an academic context, with accompanying suggestions for resolving such complications.

Keywords: Assessment; EAP; Group oral; Stance; Engagement

1) Introduction

In a country like Hong Kong (HK), where English is the medium of instruction (MOI) in tertiary education but where most students' primary first language and much of their secondary education is conducted in either Cantonese or Mandarin Chinese, the transition from secondary to tertiary education represents a considerable linguistic challenge for freshman undergraduates. Potential structural disadvantages include a recent shift in HK from three to four years of tertiary undergraduate instruction resulting in one year less of secondary education in which to prepare for academia, as well as the shift from English to Chinese as the MOI in most HK schools since the 1997 handover of sovereignty from the UK (Lo and Lo, 2014). Researchers of the kind of academic language students have to master during the first crucial months at university consider such language an ‘alien form of literacy […with] many students arriving at university thinking they have landed on Mars’ (Hyland, 2016a:246), and this appraisal is becoming increasingly applicable to HK students for the reasons outlined above. Evans and Morrison (2011) provide evidence of four key areas where students in Hong Kong experience language difficulty upon entering university, namely understanding technical vocabulary, listening to lectures, writing in an appropriate style, and conforming to the conventions of academic discourse. This paper is concerned with the latter of these issues, focusing on the assessment of academic discourse in the spoken register.

Difficulties with conforming to conventions of academic discourse are commonly addressed via freshman pre- or in-sessional courses on English for academic purposes (EAP). At the tertiary level, being able to simply converse on a topic is insufficient. Specifically within academic tutorial discussions, students are expected to formulate a *stance* on the topic of discussion, *support* their stance with examples drawn from appropriate academic sources, *defend* their stance in the face of challenges from others in the group, and *critically engage* the stance of others. Stance is 'something of a catch-all yet elusive concept' (Crosthwaite and Jiang, in press) referring to the linguistic projection of a language user's views toward the topic under discussion, while engagement involves the dialogic way in which speakers ‘relate [to their listeners] with respect to the positions advanced’ (Hyland, 2016b:169). Biber (2006) describes stance as 'attitudes that a speaker has about information, how certain they are about its veracity, how they obtained access to the information, and what perspective they are taking' (p. 87), while Hyland (2016a) describes stance as the triangulation of three important rhetorical questions that a speaker may bring to any statement:

How certain do I want to be about this?
What is my attitude towards it?
Do I want to make myself prominent here? (Hyland, 2016a, p. 248).

Mastery of a range of rhetorical strategies in the L2 is required for the successful and appropriate presentation of stance and engagement in an academic tutorial discussion setting, from the asking of direct or indirect questions to other candidates, rebuttals of other candidates' claims, the derivation of counter-arguments to a speakers' own stance, and the presentation of facts, opinions or statistics from academic sources (orally presented as ‘spoken citations’). Knowledge is also required of a variety of register-appropriate linguistic devices such as hedging, boosting, self-mentions or attitude markers used to ‘stamp their personal authority or beliefs onto their arguments’ (Hyland, 2016a:247). However, in HK, due to the time spent preparing for high school examinations in a ‘competitive exam-oriented system’ (Kennedy, 2002:439), freshman students have had relatively little opportunity to practice their development of an academic stance and to engage with others in oral production during their secondary education, at least when compared to the amount of time spent on writing and rote memorization (Kennedy, 2002; Lee, 2008). This often results in the quality of stance and engagement produced by freshman undergraduates in HK and other contexts lying in the middle of secondary and tertiary academic expectations (Matsuda & Jeffery, 2012), and such students have been found 'to use and respond to the features of stance and voice differently' to the expectations of their academic tutors (Sancho-Guinda & Hyland, 2012:2).

In response to the difficulties outlined above for HK and other similar students at English MOI tertiary institutions, the use of peer-to-peer / group oral second language assessments over the use of one-on-one oral interviews has been a long-time feature of EAP assessment in HK (since the 1990s), following the introduction of an academic group discussion segment into the HK secondary 'Use of English' exam, as a matter of government policy. There are numerous perceived benefits to a group oral approach, from the financial benefit of testing multiple participants at once, to student engagement with non-standard / non-inner circle varieties of English (Kachru, 1985; Kirkpatrick, 2007). Other benefits include positive washback, particularly in cultures like HK where the exam-oriented culture results in deficits to speaking-focused teaching and learning goals (Shohamy et al., 1986; Van Moere, 2006), as well as the empowerment of individual learners when producing language outside of privileged interviewer/interviewee relationships with 'predictable' question / response structure (van Lier, 1989; Lazeraton, 1992; Galaczi, 2004). From a formative assessment perspective, learner-to-learner interactional assessments provide suitable conditions for 'negotiation for meaning' (Long, 1985, 1996), whereby the input test-takers receive is obligatorily modified for comprehensibility wherever breakdowns in communication occur (Krashen, 1987). Here, students are active participants in 'noticing the gap(s)' (Schmidt, 1992) in theirs' or others' linguistic knowledge as evidenced by communication breakdowns, and actively repair such breakdowns via a range of conversational repair strategies including confirmation checks ('high marks?', Pica, 1987, 1996) and clarification requests (e.g. 'what did you say?'), prompted and received by students rather than interviewers that are proficient in the target language (Foster and Ohta, 2005). Such interaction leads to the enhancement of student output (Swain and Lapkin, 1995), which, in turn, results in further learner-to-learner co-construction and negotiation of knowledge (McNamara, 1997), particularly for groups of lower-level language learners who encounter gaps in conversation more frequently than higher-level learners (Gan, 2010).

Numerous studies have explored group oral assessments in terms of how students manage interactional language in such a context, and how raters perceive student performance of said features within that context. For example, higher-rated students are better able to engage with others' ideas via a range of functional language, including suggestions, (dis)agreements, explanations and challenges (Gan, 2010). Gan notes the ability to both pursue and shift the topic of talk while still making meaningful individual contributions is seen by raters as a positive aspect of student performance in a group setting (Gan, Davison and Hamp-Lyons, 2009), although Gan's studies focus on the secondary rather than the tertiary context. He and Dai (2006) looked at the L2 Chinese tertiary context for group oral assessment, namely L2 English students taking the required Chinese College English Test–Spoken English Test (CET-SET). They note that Chinese L2 students taking this test produced a significant amount of interactional language functions but that the overall range of functions was rather limited (e.g. 'agreeing', 'disagreeing'). Other, quantitative, studies on group orals have looked at how issues of shyness (Bonk and Van Moere, 2004), assertiveness (Ockey, 2009), introversion or extroversion (Berry, 2004), talkativeness (Van Moere and Kobayashi, 2004), topic (Van Moere, 2007) and different proficiency levels between interlocutors (Iwashita, 1996) each affect rater scoring of interaction in group oral performance.

However, there is less research specific to the assessment of *academic* stance and engagement in group oral assessment in tertiary, L2 contexts. Specific to group L2 oral academic assessments, raters need to negotiate the frequent real-time breakdowns in grammar and fluency expected of L2 learner production, while keeping track of the overarching points each individual speaker is trying to present, how well speakers engage with the points others are making, and whether speakers' points are presented in an appropriate academic register or are well-supported by cited evidence from appropriate academic sources. Managing these concerns simultaneously for multiple participants over extended periods is likely to result in considerable cognitive load for individual raters, leading to substantial potential for intra- and inter-rater variability on grading decisions and the practice adopted in reaching such decisions. Such concerns remain relatively underexplored in EAP assessment research.

The present study is concerned with the 'how' and 'why' rather than the 'what' or 'how much' questions of academic group oral assessments, and in answering these questions, a qualitative approach is recommended (Leung, 2012). Qualitative research on group oral tests has tended to follow a conversational analysis methodology, including Davies (2009), Galaczi (2004) and Lazaraton and Davies (2008). In Asian L2 contexts, Nakatsuhara (2011, 2013) used conversation analysis techniques on students taking oral tests in groups of three or four, analysing the data in terms of goal orientation, interactional contingency and quantitative dominance on academic performance, and noting effects of test-taker characteristics and the number of participants as influences affecting the assessment of these factors. Luk (2010) has used discourse analysis techniques to look specifically at engagement practices by HK L2 English learners (at the secondary level at least), noting that interaction among these students is contrived and designed to score points rather than contribute to any meaningful attempt at two-way communication. However, these kind of studies are focused on the test-takers' production rather than the process of rating the assessment of said production from the point of view of the raters themselves. Viewing the assessment from the perspective of the rater allows one to isolate what is salient to raters, what is difficult for them as raters, and how raters differ in their beliefs and approach as they go about the business of rating student performance in real time.

In this respect, this paper examines raters' conceptions of academic stance and engagement during a group oral EAP tutorial discussion assessment, via the collection of raters' think-aloud verbal protocols. Our purpose is to determine specific moments in the overall discussion that triggered raters' attention, caused them to make positive or negative appraisals of student performance, see when and where raters reach agreement or disagreement, and show whether raters reach similar appraisals of student performance based on different approaches, or reach differing appraisals of student performance based on similar approaches. Here, we adopt a think-aloud protocol methodology, which is outlined in detail in the following section.

2. Think-aloud studies on language assessment

Verbal reports including think-aloud protocols (TAPs) have generally been used to validate tests and to determine test-taker strategy (Bowles, 2010). However, studies using verbal reports are given frequent criticism in terms of their validity as a research methodology. Ericsson and Simon (1984/1993) in a seminal book on TAPs claim that raters think-aloud protocols suffer from issues with *veridicality* (i.e. whether raters' report their true and complete thoughts on a grading decision) and *reactivity* (i.e. whether the process of doing the protocols alters the outcome of the rating processes). Veridicality is mainly related to retrospective reports if raters could not accurately recall the procedure they are reporting on, while reactivity is related to concurrent reports if raters' practice during a rating task is affected by the performance of the verbal protocol (Bowles, 2010). In addition, TAPs are often considered difficult to administer in terms of rater training, transcribing and coding (Smagorinsky, 1994; Green, 1998). Yet, despite these shortcomings, TAPs offer an immediate and specific insight into actual – rather than perceived - rater behavior, allowing the researcher to unpack assessor's qualitative judgements on student production with reference to the circumstances that lead to such judgments (Barkaoui, 2011; Leung, 2012).

Much work has now been done on the assessment of writing using TAPs. For example, Barkaoui (2007) used TAPs on four raters with regards to the validation of a set of rating scales for L2 writing. While there were generally high levels of inter-rater agreement on the scales, the TAP data suggested that the raters themselves were the largest source of variability in terms of scores and decision making behavior. Weigle (1999) has triangulated many-facet Rasch analysis with TAP data, noting variation in how a scoring rubric could be applied to a given prompt, as well as differences in how experienced and inexperienced raters viewed the appropriacy of the prompts. Gebril & Plakans (2014) have used TAPs to understand how raters assigned grades to integrated reading / writing tasks, finding that raters only focused on surface details for low L2 proficiency texts, considering macro-structural content only for higher level texts. Leung (2012) used TAPs with four English teachers marking 12 pieces of student writing, allowing them to explore their role as teacher-assessors during the process. In the HK context, Davison (2004) used TAPs to compare cultural practices in language assessment between Australian and HK secondary school teachers, noting a 'growing concern' (p.305) about a perceived lack of understanding of teacher's decision-making processes and finding that students were often being given the same grade but for different reasons.

There have also been numerous uses of TAPs for the assessment of speaking. Wei & Llosa (2015) have used TAPs to determine differences between American and Indian raters for TOEFL, noting that the Indian raters were better able to determine the features of Indian English than the American raters, but that there were no differences between rater types in terms of scoring or severity. Kim (2015) has used TAPs with a sample of three rater's assessment of 18

ESL learners' oral performance, noting effects of rater background (including rating and teaching experience, training and educational background) on rating performance arising from the data, and how raters approach analytic rating scales. Pollitt & Murray (1996) and Hubbard, Gilbert & Pidcock (1996) have each focused on rater attention and variation in individual interpretations of student's spoken performance via TAP methodology.

Specific to group orals, Ducasse and Brown (2009) used concurrent TAPs to explore raters' conceptions of successful group oral assessment, finding that raters considered interactive listening and interactional management as constitutive of highly-rated student performance in such a context. However, concurrent TAP studies on group orals are still relatively rare, with more studies on group orals utilising retrospective verbal reports. Such studies include Orr (2002) who used retrospective verbal reports with raters assessing the Cambridge First Certificate in English, finding that raters were positive to candidates who held the floor on a range of topics and provided help to others. Likewise, in the Asian tertiary EAP context, May (2011) used retrospective verbal reports to determine which interactional features were salient to raters assessing the interactional competence of Chinese students in an EAP group discussion task, noting that co-construction of discourse via turn-taking and initiating topics was a feature associated with successful performance. May also suggests that verbal report data can be used to drive improvements to the rating scales used on such assessments, particularly when focusing on what the raters found positive or negative about students' performance in a group setting.

However, May's and others' studies on group orals reported above do not touch on many of the issues related directly to the assessment of *academic* stance and engagement, such as the appropriacy of stance in terms of an academic register, assessment of stance in real time, constructing and defending stance in the face of the stance of others, or the use of spoken citation to support an academic stance. This shortcoming leads to the need for the present study.

*3. **Current Study***

The current study was part of a larger project to validate a group oral speaking assessment in place on a tertiary EAP course using both quantitative and qualitative measures. (Crosthwaite, Boynton & Cole, 2016). This qualitative study aims to identify raters' conceptions of students' academic stance *in real time* during the assessment itself via TAP. The following questions were addressed:

RQ1) Which features of group presentation performance are salient to raters as they assess students' academic stance and engagement in a group oral assessment context?
RQ2) What difficulties do raters experience as they assess academic stance and engagement, specific to the group oral assessment context?

4. Context of the study

The assessment in question is conducted at the applied English language studies centre of a leading English MOI university in Hong Kong. In 2012, the centre developed a general EAP course comprising 36 contact and 120 learning hours ('Core University English') for all freshman students, targeting approximately 2800 students over the two semesters. Course assessment comprises two writing tasks worth 40% of the grade, out-of-class learning activities (20%), and the group oral tutorial discussion (40%). The group discussion requires students do background reading on a topic supplied 72 hours prior, making bullet-point notes and listing their academic sources on a single A4 page-sized template that students may use during the assessment. Students discuss a topic in groups of five (or fewer if a student is sick/absent). The discussions are video-recorded.

Raters are trained to mark this assessment via online standardization. They watch a full video-recorded performance, rating as they go along, then enter the final grades for each student into an online portal. Raters always rate students that were taught by other colleagues. Teachers may watch the videos more than once, but due to time constraints, the majority of teachers watch each video only once. Moderation / standardization involves viewing a single video from a past semester and grading each student, then meeting to discuss grading decisions with the course coordinator, but teachers are not asked to re-do the moderation if they were over/under the model. Individual student performance is graded, rather than the group's production as a whole.

Raters assign scores for stance using a can-do assessment scale developed by the course co-ordinators (Appendix A), under the criteria *ability to explain academic concepts and stance,* and the score for this criteria is worth 40% of the total grade alongside scores for *ability to interact with others* (30%) and *ability to communicate comprehensibly and fluently* (30%). This scale is a 12-point letter scale from A+ (highest) to D- (lowest) or F (fail). The main criteria for stance is divided into a number of sub-criteria that need to be taken into account when determining the overall grade (A+ to F), although there is no weighting of the sub-criteria in the assessment scale. For stance, these sub-criteria include the '*ability to explain academic concepts*', the '*ability to argue for a critical stance with the support of valid academic sources where appropriate*', and the '*ability to critically respond to / question other students' stance*'.

5. Method

Six teachers of the EAP course in question were recruited as raters for the purposes of this study. Raters 1, 3, 4 and 5 were monolingual speakers of English and raters 2 and 6 were bilingual speakers of Cantonese and English. Raters 1,4 and 5 were male with 2, 3 and 6 female. Ages ranged from 32 (Rater 2) to 59 (Rater 4). All raters had over ten years' teaching experience and had at least 1.5 years' experience each of teaching this particular EAP course.

The TAP methodology is as follows, similar in procedure to that of May (2011) but for a concurrent rather than retrospective verbal report. As a training session, raters were asked to provide a TAP on a 2-minute recorded segment typical of an authentic group tutorial discussion. The raters watched the video in a soundproofed room, wearing noise-reducing headphones. Sound levels were adjusted for each teacher until they felt comfortable speaking as the video was being played, in terms of both hearing their own voice over the recording, and hearing the voices on the recording over the sound of their own voice. Following this task and re-adjustment of the sound levels where necessary, teachers were asked to provide a TAP commentary on a full 25-minute group oral discussion (not the same one as seen in the 2-min segment). The 5 examinees (3 males, 2 females) were all undergraduate freshmen taking their final speaking assessment for the course. Raters were asked to verbalise their process of appraising student performance in real-time as they watched the recording. On occasion, certain teachers sometimes paused the video recording to elaborate on their comments, while others continued talking throughout. Raters were reminded that if they paused in their commentary for some time, they would be reminded to continue verbalising. As they watched, raters could make notes on pen and paper. The assessment scale was placed in front of the rater on the desk so that they could glance at the scale if need be, and could make ongoing notes as to their perceived grade under the criteria '*Ability to explain academic concepts and argue for a stance supporting by sources*' as well as the criteria for *interaction* and *comprehensibility/fluency* features as the assessment continued.

All six raters fully completed the task and left a set of suggested grades for each student with the researcher. Some raters left a summary statement commenting on the reasons for the grades they gave shortly after completing the TAP.

To avoid criticisms of veridicality and reactivity under a TAP methodology, during the TAPs themselves, raters were asked to provide comments *per se*, i.e. not to attempt to explain or justify their comments, substantially increasing the non-reactive nature of the reports as a truer representation of the raters' process in action (Ericsson and Simon, 1984/1993; Bowles, 2010). In addition, Barkaoui (2011) suggests that TAP studies should be combined with other data collection methods. To address this concern (and still following the methodology of May, 2011), participants were interviewed a week later after first watching a video recording of their TAP session, and were asked to elaborate on their TAP performance, constituting a retrospective verbal report. This session was an opportunity for raters to better explain and justify their TAP reactions, with the re-watching of the TAP video recording significantly reducing the veridicality of this interview data.

The TAPs and interview data were transcribed by a research assistant and coded together with the researcher, a monolingual speaker of English. Using MAXQDA software, we coded any segment of talk that related to student performance on stance, as well as specific reference to grading decisions, group interaction, the test itself, comments specific to language issues, and the raters' personal beliefs. 1320 comments were coded across the 6 raters across the TAP and interview data. Following coding, two additional monolingual speakers of English checked all codings for accuracy (correct/incorrect), reaching an Intraclass Correlation Coefficient of .768. Coefficient scores of between .600 and .740 are considered 'good', while scores of > .750 are considered 'excellent' (Fleiss, 1981). The monolingual English speakers met with the first author to discuss incorrect codings and these were amended when agreement was reached.

6. Results

6.1. Rater scores

Table 1 describes the letter grades assigned to each of the students for each sub-criteria of the can-do scale for stance following the TAP. For reasons of space, grades for the interaction and comprehensibility/fluency features of the assessment criteria are not shown. Each letter corresponds to each rater respectively. Grades for individual students where at least one of the raters is more than a full letter grade out (including sub-grade e.g. + or -) from another rater are highlighted in bold text.

Table 1 – Description of rater scores for stance and engagement following TAP session

Can-do statement	**Student 1** **R1 2 3 4 5 6**	**Student 2** **R1 2 3 4 5 6**	**Student 3** **R1 2 3 4 5 6**	**Student 4** **R1 2 3 4 5 6**	**Student 5** **R1 2 3 4 5 6**
Ability to express academic concepts and argue for a stance supported with sources					
Clearly explains academic concepts	**B C C B+B+C**	B+B-B B B B	**B B C B B C+**	**A-B-C C B-B**	A-A-B B+A-A-
Argue for a critical stance with sources	**B+B B B+B+C**	**A-C B B B B+**	B B B B B B-	**B+B C C B B**	B+A-B B+A-A-
Critically respond to others	**A-C C A-A-B+**	**A B B B B B+**	B B B B B B	**B+B-C C B B**	B+A-B B+A-A

From the table, it is apparent that raters appear to have varied in their assessment of stance and engagement for each student, with the exception of Student 5 where agreement was clearer. Student 5 had the most talking time of any student in the assessment (approx. 6 minutes out of 25) with the longest turns (n=5), and received higher grades overall under the comprehensibility/fluency criterion. For Students 1, 2, and 4, agreement between individual raters was generally poor, with some raters giving 'A' grades while others gave grades as low as 'C' for certain sub-criteria. Obviously, while the potential reactivity involved in raters' ability to fairly grade student performance during TAPs may account for much of the difference in grading decisions (and so the grades in Table 1 are for reference only and cannot be used to make claims about real agreement), the value of TAPs is that of a window into the internal processes that may have influenced these decisions, to which we will now turn.

6.2 .Qualitative findings

Upon analyzing the coded comments from the TAP and interview data, four main themes emerged in terms of rater conceptions of academic stance and engagement. These included student management of the register-appropriate language features involved in stance and engagement, supporting stance with academic sources, shifts in rater attention during the assessment and concerns related to assessing stance and engagement specific to a group context. We now address each theme in turn.

6.2.1. Raters' conceptions of stance and engagement

In terms of raters' conception of successful academic stance, a key issue frequently raised by raters was that of the appropriacy of language features involved in an appropriate academic stance. The majority of comments here are related to the issue of the teacher-as-rater, in that as the raters were also experienced teachers of the EAP course in question, they were often predisposed to listening out for the specific linguistic stance features that they had taught on the course, such as hedging or boosting devices. In this TAP excerpt, Rater 1 comments on this issue after hearing the use of the hedge ‘to a large extent’ during a student’s talk:

> (R1-11:40-TAP) *This is something I pick up a lot on, you get these key phrases that you may have taught on the course, so things like ‘to a large extent’ where it is obvious that this is something we taught on the course, you know to hedge claims where appropriate so when I hear the expressions that the students have been taught as part of the course this is just something that ‘activates’ for me, and then I’ll be more positive that students if I hear things that have come out on the course.*

Another major theme arising from the TAPs was the relationship between successful presentation of stance, and the comprehensibility and fluency of the L2 production involved in said presentation. While raters are supposed to grade these concerns separately according to the can-do scales they have been using in this context, it is apparent from the comments that student inability to produce comprehensible and fluent discourse means that students are being penalised twice by raters. This finding was also a feature of an exploratory factor analysis performed on the same test in Crosthwaite, Boynton & Cole (2016), with factor loadings for the criteria ‘ability to explain academic concepts’ and the three subcriteria for language concerns appearing along the same factorial dimension in that study. In the following excerpt, Student 3’s fluency is poor in this (his first) turn. Raters 1, 3, 4 and 5 each comment on the students’ difficulty with comprehensibility and fluency, yet are at the same time making appraisals of the students’ presentation of stance. Sections of individual comments where this is the case have been highlighted in bold italics (Table 2):

Table 2. Double penalization of Stance and Language.

(R1) struggling to find the words to say what he wants to say a bit (R3) some grammar mistakes, ***some really interfere with his point*** (R5) He’s gone now back to his notes. He’s doing quite a bit of reading again now. (R5) In terms of ability to communicate he’s a little bit halting in his delivery which ***makes it a little bit hard to follow what he’s saying.*** (R1) so I think that his stance is quite clear and ***if he was more fluent I think it would be a very successful turn***	Student 3: Yea, I think we've so far discuss many, um, benefit, ah, derived from the economic aspect and preserving the environment, but i think as students mentioned, ah, dealing with the to what extent question, we can, ah, we could state, ah, another alternative to compare with ecotourism and maybe I will discuss about the ,ah, the education aspect, ah, to compare with the ecotourism. And for education in Taiwan, they, ah, they promote a terms call local knowledge that is, ah, giving, ah, ah, giving (.2) knowledge of preserving environment in Taiwan, this is stated in the book called, ah, Tourism, Ecotourism and Protecting Areas, ah, written in 1996, and Taiwan is, ah, ah, having this, um, policy that to educate their citizens to preserve the environment, but compare with the ecotourism, I think, ah, this may not be better than ecotourism because ecotourism, ah, as you've mention it, provide economic profit and also it pre- preserve (.1) the, en- environment and also it provides a opportunity to educate the, ah, tourist as well, but this is, um, local knowledge and cannot satisfy, because, ah, local knowledge, ah, just only (.1) educate the citizens but not the tourist, so i think, ah […]	(R4) Ok, a little bit ready to language discuss about having this policy so maybe language ***is sort of making a little bit difficult for him to express himself*** (R4) What I'm tending to do here again, I'm speaking generally I'm tending to focus more on the person's spoken ability and ***it's taking away from my ability to really follow what they are talking about***

This finding is potentially very serious for the test-takers, given that, in the can-do scale, stance and language issues are worth a combined 70% of the total grade. In another excerpt, Rater 3 appears to be in total disagreement regarding the overall presentation of stance with the other raters during their TAP performance in terms of the specific, formulaic nature of the language used, and the tone of its delivery (R3 gave this student much lower grades than the other raters under the criteria 'ability to explain academic concepts'). Here, the difference between Rater 3 and the other raters’ appraisal of the Student 5’s stance appears to be sourced in the personal belief that the students’ language is ‘unnatural’ or ‘memorised’ (Table 3).

Table 3. Further disagreement on academic stance

(R1) She is setting up the discussion quite well (R1) She is also using some hedging with phrases like 'to some extent' (R2) She also try to state her stance quite clearly at the beginning (R4) Ok explain academic concepts well (R2) [S5] is able to explain the concepts (R5) she seems to be able to clearly explain academic concepts. (R4) student 5 completely explains academic concepts	Student 5: I think ecotourism should be involved to a moderate extent protecting ecological important areas. To answer the question of to what extent, I suggest we first compare ecotourism with other alternatives in this areas. I think one is the conventional tourism, the other is probably no tourism, or just limited to a very small extent. So, according to a book written by David in 2007, United States Ecotourism Society define ecotourism as responsible travel to specially undisturbed natural areas that conserve the environment and promote local well-being. So, first for comparing with the conventional ones, ecotourism has obvious advantages in the las- in that it lays more emphasis on protecting the environment and making profit at the same time […]	(R3) **But it doesn't seem very natural**, (R3)**it seems like she's perhaps memorize some of that** (R3) **but it doesn't sound too natural** (R3) **it sounds a little bit like a presentation or a speech** (R3) but I think there are students seems to be having to listen to her quite carefully because of the production itself, seems a little bit too much

The data in this section suggests that raters' appraisal of students' academic stance - a top-down, meaning-focused concept - is strongly affected by the bottom-up linguistic features involved in stance construction, and for the test-takers, it is difficult to have one without the other, and are punished for both when errors or disfluencies occur.

6.2.2. Supporting stance with sources

During the TAPs, numerous disagreements were sourced over the use of appropriate academic sources to support their stance, mostly in the appropriacy of the 'spoken citation' format used by test-takers (this format is taught in a number of lessons on the EAP course). In the following excerpt, two raters disagree on the quality of Student 2's use of sources, with Rater 1 suggesting the format of the spoken citation was inappropriate, while Rater 2 considers the use of sources here successful (Table 4):

Table 4 - Disagreement on citation

(R2) that's good, **we are seeing use of sources here.** (R2) the expression like '*from this evidence we can see*' that shows that **he tries to use the evidence to support his stance**	Student 2: So, according to a book Natural Area Ecotourism, Ecolo - Ecology, Impacts and Management by David, Susan and Rose in 2013, they stated that at least 40 percent of pe- of the people in the local community will be benefited economically from ecotourism, because main mainly for the increased income for the- from the job oppor- opportunities. Therefore, from this evidence, we can see that actually ec- ecotourism is a sustainable way to protect their environment and also develop that area s- so that the local community can al- also benefit economically.	(R1) and he is giving **too much information in a spoken citation** including the title the authors' name for this he had to look at the notes for quite a while

Issues with raters' conception of the use of citation and sources for stance were also noted during the follow-up interviews. Here, Rater 1 comments on the overall usefulness of the sources the students bring with them to the test, while Rater 5 is concerned with the validity of the source, given that the raters only see the reference to the source on a handout, and have likely not read the source itself:

> (Rater 1-Interview) *We are listening out for sources, and students who don't use them get penalised, but if we are forcing them to use sources and not all the sources are that useful, we are crediting them, but at the same time we are not always penalising them for sources that aren't that useful. Generally, if someone uses a source we generally give them a good score, but then are we also giving them a bad score for explaining academic concepts if the source is not useful?*
>
> (Rater 5-Interview) *So, it does seem a bit repetitive in terms of use of the source, I don't know whether it is support from valid academic sources, whether it's just one source, etc. I don't know whether that's a problem, if it's not in the criteria.*

6.2.3. Rater attention and the assessment of stance and engagement

Raters frequently commented on difficulties related to the assessment of stance and engagement over extended sequences of discourse, explicitly mentioning what they were paying attention to as a rater and how they were reading the situation in the videos in real time as the discussion progresses. In these two excerpts, Raters 1 and 4 make comments on their conceptions of a student's stance at almost identical points in the video, yet the order in which they focus on particular features of the student's performance is reversed:

> (Rater 1-20:50-TAP). *This is where the flow that I get into and what I am looking for its almost like it occurs in sequence, so I generally start with linking the turn, then I'll listen out for what his stance is in the first few sentences, then I'll check for stuff we have taught on the course metadiscourse wise, so things like signposting. Then I'll start to make judgements on grammar, after listening for a while, then as the turn continues get into fluency.*
>
> (R4-20:20-TAP) *I got to the stage now where I know pretty well, I'm not getting distracted anymore by the - I'm not focusing at all on their language ability now, now I really just need to focus more on the totality of what they are saying, and occasionally on things like linkages and citation that they use.*

This shifting of attention between students' overall academic argument and the linguistic features that constitute said argument is therefore potentially responsible for some of the variation between these raters in terms of the markedly different final grades they awarded under this criteria. Another key concern was related to the cognitive load involved in tracking students' appropriate presentation of stance over the 25 minute duration of the assessment. In this excerpt, Rater 1 comments specifically on his inability to assess student stance given the complexity of the process overall:

> (R4-41:49- TAP) *For stuff like criteria 1, explain concepts and argue stance, [...] there are so many factors to consider [...] so difficult to be able to follow a person's argument, there are so many things going through my mind...*

Moreover, many raters commented on the scalar descriptor they were to use for the stance/engagement criterion, with the scales labelled as 'always' ('A' grades) and 'almost always' ('B' grades). Here, Rater 1 mentions the difficulty involved with this process during the follow-up interview session:

> (R1-Interview) *We have these scalar options on our criteria…I've given a lot of B's for grammar here, but when you look at the criteria they have to 'always' do it to get an A, but, if they don't do it once are they not 'always' doing it? So, it's difficult when you do this because you are confused about whether they always did something [or not]. I mean, over 25 minutes you can find fault with every student in some way, so then does that mean everyone gets a B?*

The data outlined in this section thus suggest a significant effect of rater attention on the conception of academic stance in real time over the lengthy duration of the oral assessment.

6.2.4. Performance of stance and engagement in groups

The majority of comments on the group aspect of this oral assessment were sourced in the follow-up interviews. During the TAPs, most comments along this theme were related to engagement in particular, namely how students linked their turns to what others' have said, and whether a students' production constituted a critical engagement with the stance of others. As shown in section 6.2.3, raters' attention shifts considerably during the complete oral assessment, and this also potentially impacted rater conception of students' engagement with others. In this excerpt, Raters 1 and 5 make their appraisal of student response at the beginning of Student 3's turn, while Rater 6 waits until near the end, with significant variation in the final appraisal made (Table 5):

Table 5. Disagreement on linking turns.

(R5) This is now Speaker 3. He's the first one I think to ask some kind of question based on what he's heard before. **I think also this shows pretty good ability to interact**	Student 3: Um I have a question about the conflict just like ah we've mention just now the conflict between the economic profit and ah ah preserving environment by ecotourism and i think it it may not be a big deal because it is ah mostly ah depends on the rationale behind the government's ah policy[…] protecting the ecology and many of these examples are they couldn't generate economic profits from that just that we've mention maybe ah by education or by any other method most likely they couldn't generate ah income from this policy but the government is still willing to implement this is because they want to preserve the environment so I think the conflict ah harassment maybe called tourism is because the government is too concern about the economic profit but not the ah preserving the environment.	(R1) he is now bringing in more evidence use of another example and then the phrases like 'just like we mention' **so he is linking his turn to what has been said before** (R6) I don't, yeah actually he talked about something, he said, he talked about something, and he made a point but **I don't quite understand how his, how he can relate back to his stance and how it can link to the things that have been discussed by the previous speakers.**

During the follow-up interviews, many raters commented on their grading strategy, specifically for group assignments. In the moderation meetings, raters are instructed to judge each students' individual performance against the can-do scales, yet the following two excerpts show that certain raters bring their own strategy to this process. Rater 4's comments suggest that they assign every student coming into the test a default ‘B’ grade, measuring individual student performance from that baseline over time, and in comparison with others:

> (R4-Interview) *The other S2, S3 and S4, I just gave a full B for all, I couldn't really, nothing really stood out*

Rater 5 preferred to make more specific comparisons between test-takers, using the best performing student as a baseline from which to compare the performance of the others:

> (R5-Interview) *I also compare each one as well, you know I sort of had a grading base on each one and of course number 5 was the best, number 1 and 2 weren't so good, what I thought would be the grades, generally I start of the default B grade and then just go up and down from there.*

Specific to concepts of successful engagement in a group setting, raters often made personal comments in the interviews regarding the ‘typical’ situation of an academic group assessment, and how deviations from this ‘norm’ will trigger specific decisions on grading. Students’ facilitation of discussion between others was seen as a strong positive for Rater 1:

> (R1-Interview) *Generally with these group discussions there’s usually one character in it who tries to set up opposing viewpoints, and these people usually do better when it comes to critically responding to others.*

Rater 6 commented specifically on the general lack of direct questions with a typical L2 EAP cohort, and saw any attempt to ask such questions as a strong positive:

> (R6-Interview) *Students directly questioning other students […] we generally don’t see this kind of back and forth in the tutorial discussion practices and we very rarely see students directly questioning other students. It’s something I like to see, because it puts students on the spot and it also shows you are being critical*

Another issue related to the academic nature of the group oral was that of the usefulness of individual prior preparation in a group tutorial discussion setting. Students are given 72 hours to make notes from academic sources, but due to the unpredictable nature of the assessment format in which raters do not use rubrics to guide the discussion over the 25 minutes, it may be the case that students do not get to use the statistics or data they have read. Rater 2 comments on this point specifically during the interviews:

> *(*Rater 2 - Interview*)* *The design of this assessment makes it hard if students have prepared some ideas which are not, like, not mentioned, not discussed by other students,*

> *then they will have nothing to say. So, sometimes students cannot respond to others' points just because they may not have relevant sources to what came up in the discussion.*

Here, it is likely that the group aspect of this assessment could lead to difficulty rating student performance across different test situations (person-by-occasion), even where students have done considerable prior planning, if the topic under discussion or the nature of the interaction at a given moment doesn't allow a student to utilise the information they have spent time preparing.

7. Discussion

This paper has outlined a number of themes arising from the data in terms of how raters conceive of the academic stance and engagement produced by students in a group oral assessment context.

In terms of RQ1 (matters salient to raters in the assessment of academic stance and engagement), we have captured, via think-aloud verbal protocols and post-hoc interviews significant differences in how individual raters conceptualise and then judge student presentation of academic stance and engagement in real time. Second language comprehensibility concerns, as well as the appropriateness of spoken citation appear to be key issues for raters when conceptualising (un)successful academic stance and engagement. Raters are also sensitive to whether or not students use interactive language that teacher-raters have taught on the course, and are seemingly unable to prevent themselves, as teachers of the course, from bringing with them a set of predefined attitudes towards the 'typical' situational context of the group tutorial discussion in this L2 EAP context, with reported concomitant effects on the positive / negative conception and appraisal of student performance.

The finding that students are being penalized twice for stance and language issues is in line with the exploratory factor analysis used in our previous quantitative study on the same assessment (Crosthwaite, Boynton & Cole, 2016) as well as previous assessment research (for writing, at least, Cumming 1990; Gebril and Plakans, 2014) suggesting that language-related issues result in raters being unable to work out the overall stance inherent within a student's turn. Here, it may be beneficial to combine language and content concerns into a single criterion, and to replace to scalar judgements of performance such as 'always', 'usually' etc. to more detailed 'can do' criteria. To resolve these issues, we consider the need for teachers to work together on modifications to the can-do scale in place in the present study, so as to ensure that the considerations of the can-do scale are fully internalised by teachers, following Weigle (2002), Hamp Lyons (1991) and Barkaoui (2007).

In terms of the use of academic sources in oral assessments, one suggestion is to consider the need for one or two *shared* sources in a group academic discussion, with students also able to select two of their own. This approach would help with the unpredictability of interactive extended discourse and its effect on student preparation, and also help with rater familiarity and engagement with the source text as well. We follow Gebril & Plakans (2014) who consider this approach as essential in the grading of citation format and mechanics alongside student engagement with source texts – a key source of contention among the raters in our study – while Wiseman (2012) suggests that such engagement works to mitigate rater severity when scoring performance on academic assessments.

Moreover, it is apparent from the findings that even with a standardized rubric (or at least a supposed shared understanding of what the criteria entail), raters bring conflicting accounts of the criteria when conceptualising test-taker's production (following Barkaoui, 2007). On this occasion, they conflict on their conceptions of academic stance and engagement both at particular times in the students' turn, or across the assessment as a whole. Even with changes to the assessment scale as suggested above, we have also seen (as with Lumley, 2002; Gebril and Plakans, 2014) that raters emphasise certain features of a student's production over others (even if raters understand the criteria similarly). While this might sound like a simple 'raters vary', it highlights the call for improved training for teacher raters on the assessment of academic oral discourse, and perhaps greater consequences for raters who do not meet the required standards during standardisation activities, in a bid to reduce the impact of raters' personal beliefs and background on the rating process. Rater training is successful in reducing severity (Wiegle, 1998) and improving inter-rater correlation and agreement (Davis, 2016) across cohorts of raters, and verbal reports have even been used to describe the effect of such training on rater's interpretation of assessment criteria (Wiegle, 1994), although as Wiegle (1998) notes, the emphasis of training should not be 'to force raters into agreement with each other […] but rather to train raters to be self-consistent' (p.265). Raters should be also given more feedback on their performance during rating (Wiseman, 2012) while Davis (2016) found that ‘scoring tools’ including exemplar responses and activities designed around rubrics increase rater accuracy during training. However, rater training should be done in tandem with other methods, rather than a magic bullet. We are also mindful of the additional resources as well as damage to potential staff morale that a pass/fail (and repeat) type standardisation process would create, given that teaching staff are generally already pushed to their limits.

For RQ2 (difficulties reported by raters for group oral assessment), we have found a significant impact of rater attention on their conceptions of stance and engagement in real-time, and the unpredictable nature of group interaction on students' eventual performance. These findings raise concern over whether the raters' conceptions of stance and engagement in extended discourse under examination conditions are really a true reflection of what a student has (or has not) achieved. Specific to the concerns regarding rater attention, it would help to reduce the amount of time spent on discussion, and / or reduce the number of candidates present in any given assessment. We note that it would take the same amount of time (bar the practicalities involved with switching candidates in the assessment room) for raters to assess four students in 20 minutes as it would to assess five students over 25 minutes, with each student still having an average five minutes opportunity to present their stance and engage with others. Doing so would reduce the cognitive strain on the rater during the assessment, and reduce the number of participants’ performances the rater has to track simultaneously.

Specific to the group aspect of this assessment context, the data has also revealed that - despite not being a feature of the standardisation process - raters can and do compare individual student performance against that of other student performance in the group, leading to potential inconsistency across varying test situations with students of varying proficiency. Moreover, the data revealed that teacher raters bring with them expectations and prejudices about how students interact in group academic settings, and these appear to influence raters' conceptions of successful stance production and engagement. Numerous studies have pointed that out the characteristics of raters themselves often influence grading decisions and that such characteristics may not be stable across various test situations (Lumley and McNamara, 1995) or may vary according to individual criteria (Brown, 1995). However, our quantitative study (Crosthwaite, Boynton & Cole, 2016) on the same assessment found little effect of rater variables on the assessment of stance and engagement across multiple assessments of this particular examination. Given the qualitative focus of the present study, we do not wish to attempt to comprehensively link any variation found between the raters in this study to their quantitative variables (e.g. age, length of time teaching the course). We do, however, acknowledge that this could be considered as a potential limitation of the study. One suggestion to overcome the unpredictable nature of the group oral across test situations is to rate student production across a portfolio of sessions / topics. Currently, students have three trial discussions and one full mock speaking test before their final assessment. Assessment over numerous discussions and / or raters could promote a more accurate reflection of student performance for individuals given the potential impact of topic, preparation, rater or interaction on an individual's scores for one instance (East, 2006; Elliot, 2005; Hamp-Lyons, 2002; Rizaei and Lovorn, 2010), helping to avoid the impact of person-by-occasion-by-rater variance (Van Moere, 2006).

8. Closing comments

A potential limitation of the present study is that we are currently limited to only one kind of assessment in one university. However, as group discussion as a form of assessment is now increasingly used in many other contexts, we consider many of the findings in the present study as generalizable to other contexts, namely that raters exhibit substantial variation in their conceptions of successful stance production and academic engagement in group oral assessment settings, in different ways but often at the *same* point of a student's turn, and that the assessment of stance and engagement produced by multiple parties by one examiner over extended discourse is a cognitively demanding task. Such crucial information is not readily accessible by looking solely at quantitative data, thus it is important to consider how qualitative and quantitative approaches provide different *yet equally important* insights into rater behavior. The results and suggestions put forward in this paper should therefore be of use to EAP course coordinators in determining the process by which raters conceive academic stance and engagement in practice, and how group oral EAP tests a whole might be improved. The paper has also demonstrated the value of TAPs for researchers interested in exploring the practice, procedure and product of language assessment.

9. References

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12, 86-107. DOI: 10.1016/j.asw.2007.07.001

Barkaoui, K. (2011). Think-aloud protocols in research on essay writing: An empirical study of their veridicality and reactivity. *Language Testing,* 28(1), 51-75. DOI: 10.1177/0265532210376379

Berry, V. (2004). *A study of the interaction between individual personality differences and oral performance test facets*. Unpublished doctoral dissertation. King's College, University of London, UK.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers* Amsterdam: John Benjamins.

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

Bowles, M. A. (2010). *The Think-Aloud Controversy in Second Language Research*. London: Routledge.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12(1), 1-15.

Crosthwaite, P., Boynton, S. & Cole III, S. (2016). Validating an academic group tutorial discussion speaking test. International Journal of English Linguistics 6(4), 12-29.

Crosthwaite, P., Jiang, F.K. (to appear). Does EAP affect written L2 academic stance? A longitudinal learner corpus study. System, accepted, in press.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51.

Davies, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, 26(3), 367–396.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117-135. doi:10.1177/0265532215582282

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing*, *21*(3), 305-334.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), 423-443.

East, M. (2006). The impact of bilingual dictionaries on lexical sophistication and lexical accuracy in tests of L2 writing proficiency: A quantitative analysis. *Assessing Writing*, 11 (3), 179–197.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. New York: Peter Lang.

Ericsson KA and Simon HA (1984/93) *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Evans, S., & Morrison, B. (2011). The first term at university: implications for EAP. *English Language Teaching Journal*, 65(4), 387-397.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*, 5(1), 105-112.

Foster, P., & Ohta, A. S. (2005). Negotiation for meaning and peer assistance in second language classrooms. *Applied linguistics*, 26(3), 402-430.

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.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher-and lower-scoring students. *Language Testing*, 27(4), 585-602.

Gan, Z., Davison, C., & Hamp-Lyons, L. (2009). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics*, 30(3), 315-334.

Gebril, A. & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, 21, 56-73, DOI: 10.1016/j.asw.2014.03.002

Green, A. (1998) *Verbal Protocol Analysis in Language Testing Research: A Handbook*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). *Assessing Second Language Writing in Academic Contexts*. Ablex Publishing Corporation, NJ.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, 8 (1), 5–16.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing*, 23(3), 370-401.

Hubbard, C., Gilbert, S., & Pidcock, J. (2006). Assessment processes in speaking tests: A pilot verbal protocol study. *Research Notes*, 24, 14–19.

Hyland, K. (2016a). Writing with attitude: Conveying a stance in academic texts. In E. Hinkel (Ed.) *Teaching English Grammar to Speakers of Other Languages* (pp.246-265). London: Routledge.

Hyland, K. (2016b). A very peculiar practice. In R. Ellis (ed.) *Becoming and Being an Applied Linguist: The Life Histories of some Applied Linguists* (pp.155-175). Amsterdam, John Benjamins.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, 5(2), 51-66.

Kachru, B. (1985). Institutionalized second language varieties. In S. Greenbaum (Ed.), *The English Language Today* (pp. 211-226). London: Oxford University Press.

Kennedy, P. (2002). Learning cultures and learning styles: Myth-understandings about adult (Hong Kong) Chinese learners. *International Journal of Lifelong Education*, 21(5), 430-445.

Kim, H.J. (2015) A Qualitative Analysis of Rater Behavior on an L2 Speaking Assessment, *Language Assessment Quarterly*, 12:3, 239-261, DOI:10.1080/15434303.2015.1049353

Kirkpatrick, A. (2007). *World Englishes Paperback with Audio CD: Implications for International Communication and English Language Teaching*. Cambridge University Press.

Krashen, S. D. (1987). *Principles and Practices in Second Language Acquisition.* New York: Prentice-Hall.

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System*, 20, 373-386. http://dx.doi.org/10.1016/0346-251X(92)90047-7

Lazaraton, A., & Davies, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly*, 5(4), 313–335.

Lee, I. (2008). Understanding teachers' written feedback practices in Hong Kong secondary classrooms. *Journal of Second Language Writing*, 17(2), 69-85.

Leung, C. (2012). Qualitative research in language assessment. In *Encyclopedia of Applied Linguistics*. Wiley.

Long, M (1985). Input and Second Language Acquisition Theory. In S. M. Gass and C. G. Madden (eds) *Input and Second Language Acquisition* (pp.377-93). Rowley, M.A.: Newbury House

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W C. Ritchie & T. K. Bhatia (Eds.) *Handbook of Research on Language Acquisition. Vol. 2: Second Language Acquisition.* (pp. 413-468). New York: Academic Press.

Lo, Y. Y., & Lo, E. S. C. (2014). A meta-analysis of the effectiveness of English-medium education in Hong Kong. *Review of Educational Research*, 84(1), 47-73.

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly*, 7(1), 25-53.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters?. *Language Testing*, *19*(3), 246-276.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54-71.

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

Matsuda, P.K & Jeffery, J.V. (2012). Voice in student essays. In K. Hyland & C. Sancho-Guinda (eds.) *Stance and Academic Voice in Written Academic Genres* (pp.151-166). Hampshire: UK.

May, L.A. (2011) Interactional competence in a paired speaking test: features salient to raters. *Language Assessment Quarterly*, 8(2), pp.127-145.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, 28(4), 483-508.

Nakatsuhara, F. (2013). *The Co-construction of Conversation in Group Oral Tests*. Peter Lang.

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, 26(2), 161-186.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. System, 30(2), 143–154. doi:10.1016/S0346-251X(02)00002-7

Pica, T. (1987). Second-language acquisition, social interaction, and the classroom. *Applied linguistics*, *8*(1), 3-21.

Pica, T. (1996). Second Language Learning through Interaction: Multiple Perspectives. *Working Papers in Educational Linguistics*, 12(1), 1-22.

Pollitt, 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, UK: Cambridge University Press.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing*, *15*(1), 18-39.

Sancho-Guinda, C. & Hyland, K. (2012). Introduction: a context-sensitive approach to stance and voice. In K. Hyland & C. Sancho-Guinda (eds.) *Stance and Academic Voice in Written Academic Genres* (pp. 1-15). Hampshire: UK.

Schmidt, R. (1992). Awareness and second language acquisition. *Annual Review of Applied Linguistics, 13*, 206-226.

Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal*, 40(3), 212-220.

Smagorinsky P (1994) Think-aloud protocol analysis: Beyond the black box. In P Smagorinsky (ed.), *Speaking about writing: Reflections on research methodology* (pp. 3–19). Thousand Oaks, CA: Sage.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371-391.

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

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing,* 23(4), 411-440.

Van Moere, A. (2007). *Group oral tests: how does task affect candidate performance and test scores?* Unpublished doctoral dissertation, Lancaster University.

Van Moere, A. and Kobayashi, M. (2004). Group oral testing: does amount of output affect scores? Paper presented at the *Language Testing Forum*.

Wei, J. & Llosa, L. (2015). Investigating Differences Between American and Indian Raters in Assessing TOEFL iBT Speaking Tasks, *Language Assessment Quarterly*, 12:3, 283-304, DOI: 10.1080/15434303.2015.1037446

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287.

Weigle S.C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*, 145–178.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wiseman, C.S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, 17, 150-173 DOI: 10.1016/j.asw.2011.12.001

Crosthwaite, P., Boynton, S. & Cole, S. (2017). Exploring rater conceptions of academic stance and engagement during group tutorial discussion assessment. *English for Academic Purposes*, to appear.