An Exploratory Study of Rater Orientations on Pre-sessional Academic Conference Presentations through Verbal Report Protocols

by

Alasdair Braid

Dissertation submitted to the University of Nottingham in partial fulfillment

of the requirements for the Degree of Master of Arts

September 2016

Word Count: 14,951

Acknowledgements

I would like to thank all the participants who kindly gave me their time for the verbal reports and follow-up interviews. Their work is very stressful and, at the time, they had probably had their full share of assessing academic presentations but they were patient and helpful to me throughout. I particularly owe a debt of gratitude to the pre-sessional co-ordinator and the Pre-sessional 3 co-ordinator at the EAP Unit where I conducted my research. These two individuals were my contacts with the EAP Unit. They helped put me in contact with the students and the participants and gave me useful suggestions. Frankly, I am not sure how this dissertation would have been possible without them. I hope this dissertation is worthy of their altruism.

I would also like to thank my supervisor Martha Jones, for her helpful feedback and suggestions and for her patience during the early stages of the process.

Abstract

This exploratory study focuses on the Academic Conference Presentation (ACP), a frequent form of speaking assessment in EAP (English for Academic Purposes), but one that is under-researched from a rater's perspective. This study analyses raters' perceptions of the ACP performance from four aspects: what type of features are heeded by raters most, i.e. genre, criticality or language features, the construct-relevance of those heeded features (Fulcher, 2003), the clarity and processability of the rating scale, and the types of strategy that raters use to cope with this task.

The study uses a retrospective verbal report methodology, as exemplified in the studies of Orr (2002), Brown (2000) and May (2006). It focuses on an ACP summative assessment from an interim pre-sessional course delivered by the EAP Unit of a British university. The verbal reports were carried out with five trained raters who work in the EAP Unit and the data elicited from the verbal reports was triangulated with follow-up interviews with each rater (Dörnyei, 2007).

The research found that the raters heeded predominantly construct-relevant features of performance and attended to criticality and genre features more than language features. The rating scale was found to put a processing strain on the raters and used a substantial amount of relativistic (Knoch, 2009) wording. The research also found that several raters made impressionistic judgements of the overall score while the performance was in progress, but then checked this against the criteria, or against the details of the unfolding performance. There were differences in terms of the philosophies of two raters, with one rater rewarding positive features of performance to counterbalance negative features while another rater adhered more strictly to the rating scale. However, the rating approaches exhibited by the raters were complex and did not fit neatly into the synthetic and analytic types outlined by Pollitt and Murray (1996).

Contents

Figures

Tables

1. Introduction

In the two years that I have worked on EAP pre-sessionals, I have worried much more about my consistency as a rater of spoken assessment than I have as a rater of written assessment. This may be due to the real-time aspect of rating speaking exams. Although many speaking exams are video or audio-recorded, the necessity of assigning scores to a multitude of students in a short time usually prevents the rater from reviewing aspects of the performance. Consequently, raters have to make scoring decisions either on the spot or in the few minutes between one performance and the next (Weir, 2005). As a result, I was interested in reading research of raters' orientations to speaking exams to see whether the researchers found the kind of subjectivity that I felt prone to in my own rating. If we define subjectivity in terms of raters attending to non-criterion features of performance, i.e. those features that are not in the scoring rubric, then the research *did* show that trained raters often diverged from the rating scale in their evaluations of student performance (Orr, 2002; Brown, 2000; May 2006). The research also showed that raters brought different foci or 'individual frames of reference' (May, 2006, p.34) to the rating experience. For example, in May's study of two trained EAP raters, one rater focused their attention on accuracy features and the other on fluency (Ibid). In the same study, the two raters saw students' use of intertextuality in different lights. One saw it as detracting from fluency; the other saw it as showing an ability to synthesise sources (Ibid). I wanted to investigate whether some of the same idiosyncrasies could be found in trained raters of an academic genre that is commonly tested in EAP: the academic conference presentation (ACP). This genre has been studied from a student's perspective (Zappa-Holman, 2007) and from a discourse analytic perspective (particularly in the collection of papers edited by Ventola, Shalom and Thompson, 2002), but, to my knowledge, it hasn't been studied from a rater's perspective. I wanted to apply to the ACP genre the same verbal report methodology that Orr (2002), Brown (2000) and May (2006) had applied to the paired speaking (Orr and May) and oral interview tasks (Brown).

The setting of the research is an EAP Unit of a British University. The EAP Unit provides four main pre-sessional courses which students join according to their IELTS score. These courses follow a developmental path from Pre-sessional 1 at the lower levels, to Pre-sessional 4, after which students should be ready to join their departments. The focus of the present study is Pre-sessional 3 and the summative assessment for this course: that is, the test of what students have grasped during the year (Brown, 2010). This assessment is divided into two parts. The first part is a research paper written on a topic of the student's choice from their subject specialism. The topic should be linked to the theme of the course, which this year is power. The main points from this research paper, or one particular area of interest in the research paper, are then expanded into a twenty minute presentation.

The main focus of my research will be verbal reports with five trained raters from the EAP Unit's staff, through which I will analyse three main areas: the 'criterion-ness' of their rating, whether they focus on genre, criticality or linguistic aspects of the ACP performance, and how they interpret the rating scale. From my investigation of these areas, I hope to draw implications for rating scale design which I hope will be useful for a tutor from the EAP Unit. As can be seen in Fulcher's schematic, rating scale design, rater training and rater characteristics are closely intertwined in the assessment process and these three strands will run through my literature review, results and discussion (Figure 1).

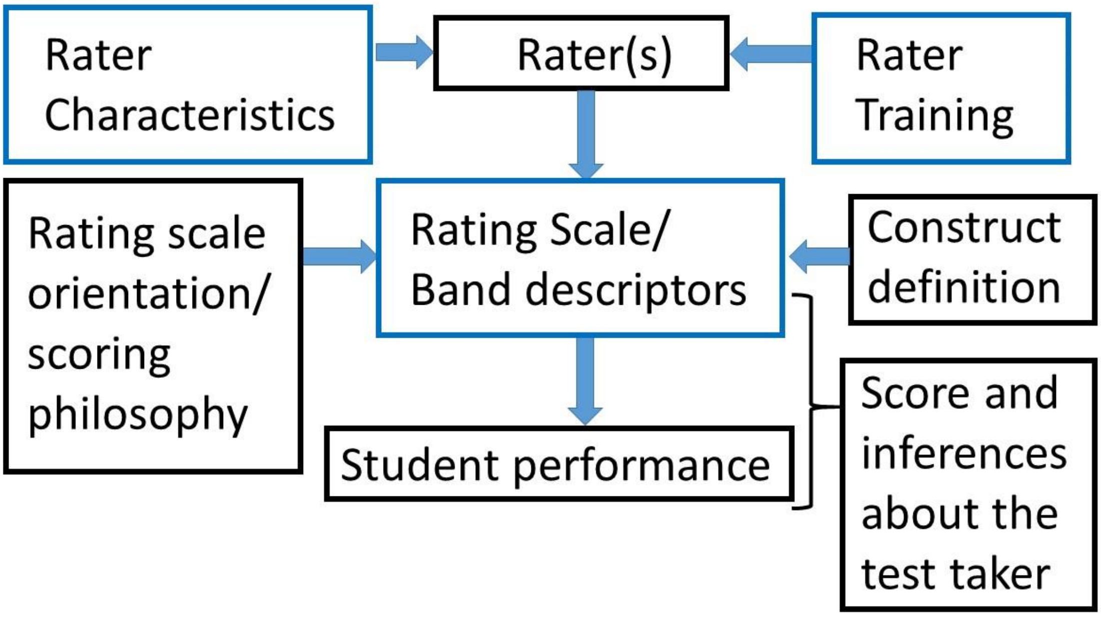

Rater Characteristics
Rater(s)
Rater Training
Rating scale orientation/ scoring philosophy
Rating Scale/ Band descriptors
Construct definition
Student performance
Score and inferences about the test taker

Figure 1: Adapted from Fulcher's (2003) Model of Speaking Test Performance (p.115).

Another thread to my research will be to investigate what strategies raters use to assess the presentations. This should shed further light on the manageability of the rating scale, as well as give useful information to rater trainers, and inexperienced raters like myself, about the kinds of strategies experienced raters use to mediate this multi-semiotic speech event (Yang, 2014a).

2. Literature Review

In the first part of the literature review, I would like to look at some discourse analyses of the ACP genre and some theories of critical thinking which will give me a language to describe the ACP performance in the results section. In the second half, I will discuss research into more practical aspects of speaking assessment such as moderation, rater training and rating scale design. Finally, I will look at some studies of rater orientations to speaking assessment and studies of rater strategies which have informed my research design. First, however, I would like to briefly explain the concept of 'construct', as it is fundamental to everything that I will talk about.

2.1. The Construct of the Test: the Academic Conference Presentation

A construct is defined by Richards and Schmitt (2010) as a 'concept that is not observed directly but is inferred on the basis of observable phenomena' (p.122). In the case of speaking tests the observable phenomena are the behaviours exhibited by the test-taker while enacting the speaking task. This allows the examiner to make an inference about the test-taker's speaking ability in general (an ability construct) or the test-taker's ability to perform a similar task in the real-world (a performance construct) (Chapelle and Brindley, 2010). The construct validity of a test, therefore, is how well the test items reflect the theoretical underpinnings, or the construct, of the test (Richards and Schmitt, 2010).

As illustrated in Fulcher's (2003) expanded model of speaking test performance (Figure 2), defining the construct occupies a central position in the test development process, informing 'the mode of delivery, tasks and method of scoring' (Ibid, p.19). The close link between the rating scale and the test construct is particularly salient, as the criteria of the rating scale are operationalisations of the test construct (Ibid). This means that rating scale development needs to be a carefully thought-through process.

For most EAP assessment, the construct is a ‘performance construct’ in that it allows examiners to make an inference ‘”directly” from test performance to performance outside the test setting’ (Chapelle and Brindley, 2010): in this case, the student’s ability to perform academic presentations in their future discipline. However, the rating scale or band descriptors used by the EAP Unit (Appendix A) contain a range of criteria which focus on language proficiency and criticality, as well as task-based criteria. These are represented by three descriptor categories: genre, criticality and language. These criteria allow a rater to make inferences about the student’s general speaking ability, their level of critical thinking as well as their performance ability. Therefore, the construct being analysed in this study could be described as a hybrid ability-performance construct (Ibid).

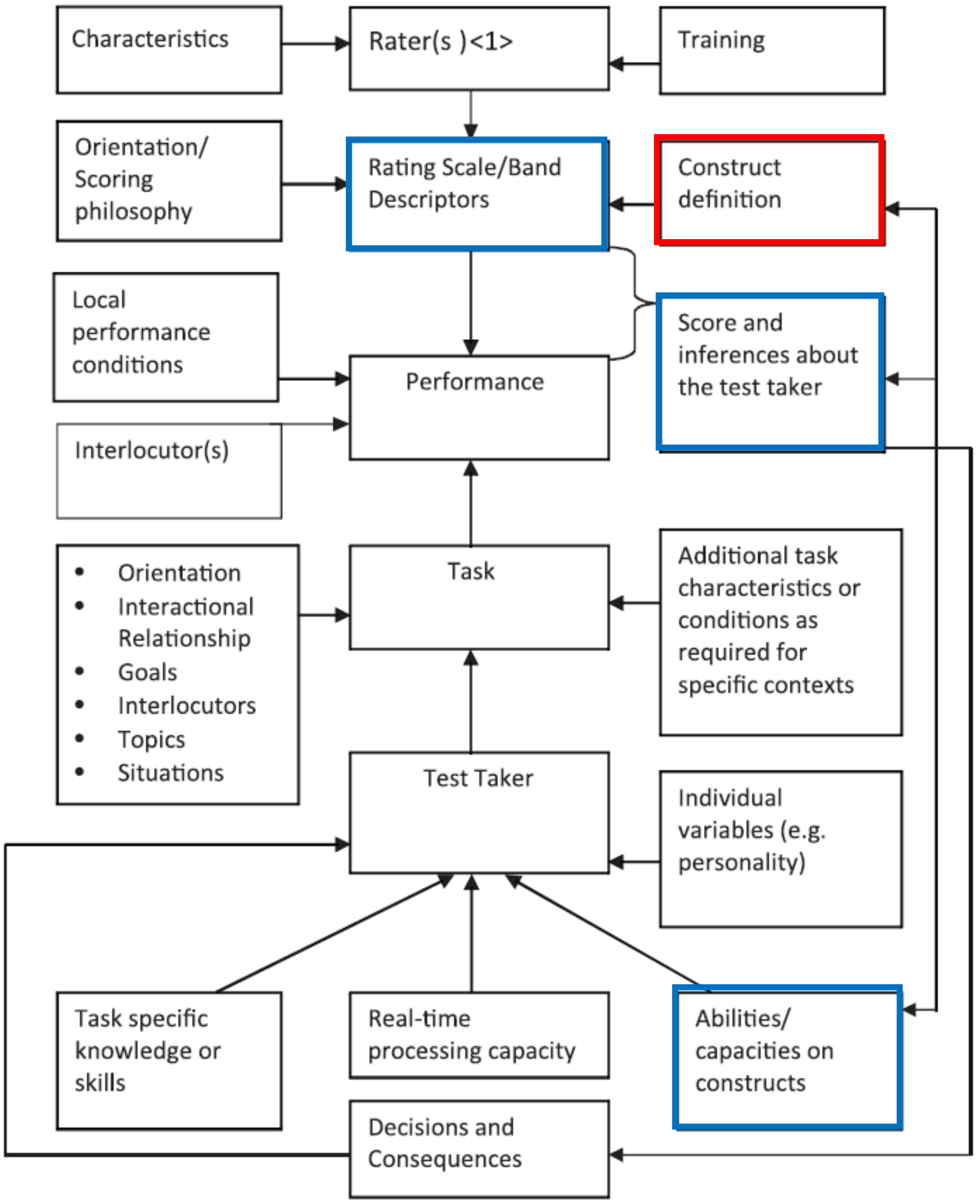

Characteristics
Rater(s)<1>
Training
Orientation/ Scoring philosophy
Rating Scale/Band Descriptors
Construct definition
Local performance conditions
Score and inferences about the test taker
Performance
Interlocutor(s)
Orientation
Interactional Relationship
Goals
Interlocutors
Topics
Situations
Task
Additional task characteristics or conditions as required for specific contexts
Test Taker
Individual variables (e.g. personality)
Task specific knowledge or skills
Real-time processing capacity
Abilities/ capacities on constructs
Decisions and Consequences

Figure 2: Fulcher's expanded model of speaking test performance (2003, p.115)

2.2. Discourse Analytical Studies of ACPs

Many researchers have drawn on Systemic Functional Linguistics (SFL) as a framework to analyse the ACP genre (Frobert-Adamo, 2002; Forey and Feng, 2016; Hood and Forey, 2005; Morell, 2015; Cassily and Ventola, 2002; Shalom, 2002). In particular, they invoke the three functions of language proposed by Halliday (1970): the ideational function, which is the way language ‘gives structure to experience, and helps to determine our way of looking at things’ (Ibid, p.143); the interpersonal function, which ‘establishes and maintains social relations’ (Ibid), and the textual function, which connects passages of discourse and creates texts (Ibid). Morell (2015) equates these to the multi-modal features of the ACP. She links the ideational function to the specific content of the presentation, the textual function to how the presentation is organised and the interpersonal function to how the speaker shows their attitude towards the topic and the audience (Ibid). Morell draws on this theory to create her four modes of presentation: the spoken, the written, the non-verbal materials (NVM) and the body language modes (Ibid). She argues that successful presentations usually include all four modes, as well as concise ideational information, organisational textual features such as discourse markers and inclusive interpersonal devices, such as greetings or attention-grabbing techniques (Ibid). These semiotic modes combine to allow the presenter to connect to the audience and communicate their intended message (Ibid).

Halliday and Hasan (1989) have described three main variables which describe the context of a speech event: the field, tenor and mode of discourse. The tenor is particularly salient to many studies of ACPs. The tenor describes who the interlocutors are, their closeness or distance, as well as ‘their statuses and roles’ (Ibid, p.12). This can have important implications for the register of the speech event (Ventola, 2002). It is particularly salient at the beginning of the conference presentation, a stage which Ventola names the contextualisation stage (Ibid) (Figure 3). So, for example, in this stage a new researcher might downplay the significance of her findings because amongst her audience there may be her peers and those academics she looks up to (Shalom, 2002). For other speakers, the tenor of the communication can be expressed by diverse discourse strategies, such as dissolving tension, showing solidarity with the audience, accommodating the audience as part of the discourse community or establishing the speaker’s status as an expert in the field (Hood and Forey, 2005).

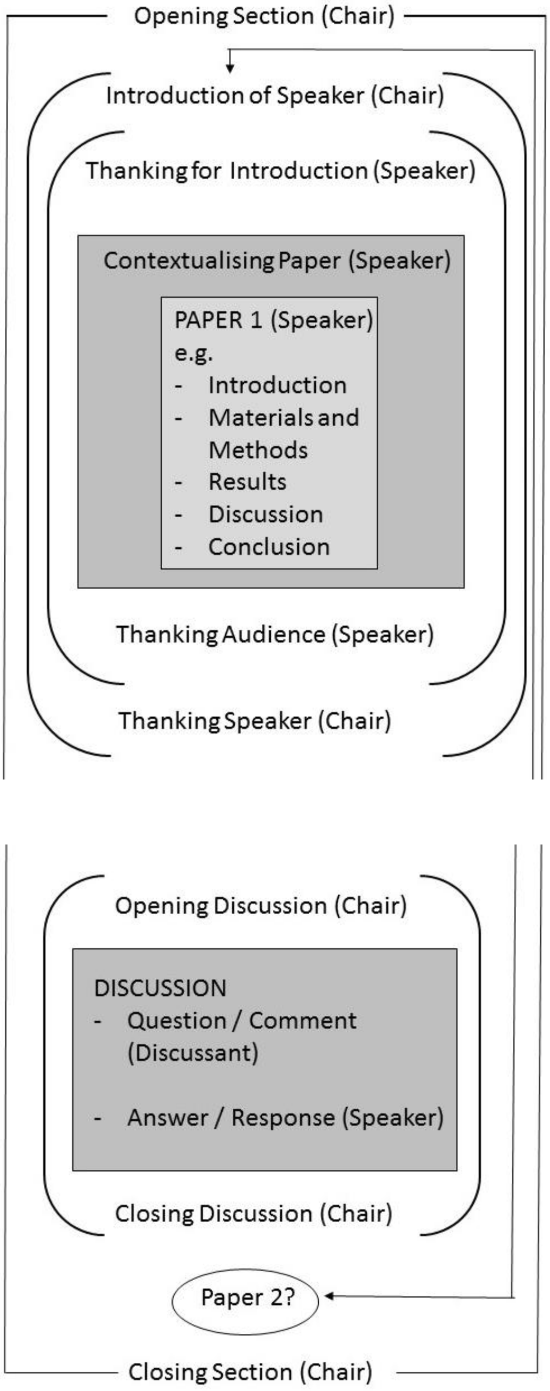

Opening Section (Chair)
Introduction of Speaker (Chair)
Thanking for Introduction (Speaker)
Contextualising Paper (Speaker)
PAPER 1 (Speaker)
e.g.
- Introduction
- Materials and Methods
- Results
- Discussion
- Conclusion
Thanking Audience (Speaker)
Thanking Speaker (Chair)
Opening Discussion (Chair)
DISCUSSION
- Question / Comment (Discussant)
- Answer / Response (Speaker)
Closing Discussion (Chair)
Paper 2?
Closing Section (Chair)

Figure 3: A generic structure of an ACP (Ventola, 2002, p.30)

The interpersonal function is also key to Morell's (2015) body language mode. Morell divides this mode into kinesics, i.e. eye contact, facial expressions and gesture, and proxemics, i.e.

physical closeness or distance to the audience and movements in space (Ibid). For Morell, kinesics and proxemics serve an important interpersonal function as well as conveying authorial voice: they 'create a specific rapport with the audience, express [the speaker's] attitude towards what, and to whom they are communicating, as well as intensify their evaluative stance' (Ibid, p.141).

I'd finally like to talk about visual support as this will be another criterion that will be discussed in my results. Morell distinguishes between three types of visual support: decorative, where the visual image simply serves a cosmetic function; illustrative, where the image contextualises what the speaker is talking about, and expository, where the image has an evidence-providing role, such as a statistical chart (Ibid). Cassily and Ventola (2002) point out two key steps or 'moves' for communicating visual information. The first move is identification. This describes the process where the speaker switches between verbal and visual modes of presentation (Ibid). In this stage, the speaker identifies the image with participants and processes in the topic 'which have been established previously through language alone' (Ibid, p.176). The second step is to contextualise the image. Whereas the identification stage makes a link between the verbal and visual information, the contextualisation stage makes a link between the image and the overall topic of the presentation and the 'academic culture of which it is part' (Ibid, p.178). This can be done through making evaluative assessments or justifying the speaker's authority to show the image (Ibid). According to Cassily and Ventola, these two discourse moves are closely intertwined in the ACP speech event but may follow different sequences (Ibid).

2.3. Critical Thinking

The BALEAP competencies, a framework for EAP best practice, make it explicit that critical thinking should be integral to EAP assessment, providing students with 'knowledge transforming tasks and activities' (2008, p.6). Defining critical thinking is difficult however, without giving long lists of critical thinking elements for each academic activity. Cottrell (2011) sums it up simply as ‘higher order thinking’. This is demonstrated neatly by one of the early theorizations of critical thinking: Bloom’s taxonomy of the cognitive domain (Bloom et al, 1956, in De Chazal, 2013) (Figure 4). In Bloom’s taxonomy we can see a progression from simply acquiring new information at the bottom of the process, to deconstructing this information into its component parts in the higher cognitive domains, synthesising it with other sources of information, and, finally, reflecting on the reliability and effectiveness of the information (De Chazal, 2013). At the top of the scale is evaluation, which De Chazal (Ibid) and Cotton (2010) link closely to authorial voice. This close relationship is apt, as the author’s choice, arrangement and critique of the different sources they use projects the author’s voice, even without the use of ‘I think’, or ‘In my opinion’ (De Chazal, 2013).

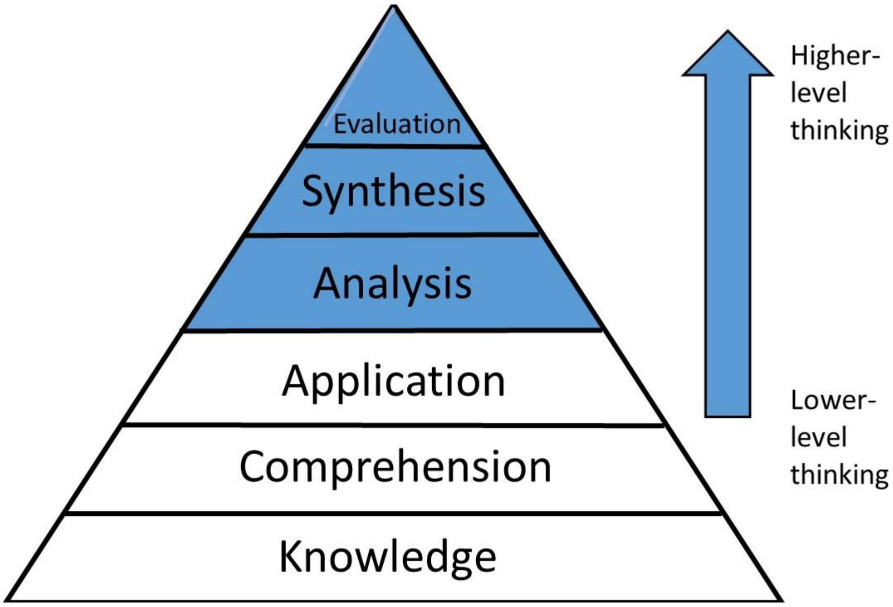


Figure 4: Bloom’s Taxonomy of the Cognitive Domain (Bloom et al, 1956, in De Chazal, 2013)

Stance has been frequently studied in the ACP genre, particularly in the work of Yang (2014a; 2014b) and Zareva (2013). Zareva focuses on the use of the personal pronoun, or the self-mention, in the ACP genre. Using the framework of Tang and John (1999), she categorises self-mentions into a continuum between the less assertive, less face-threatening self-mention and the more powerful projection of identity, which carries claims of expert knowledge. For example, at the less assertive end of the spectrum is the 'I' as a guide who leads the listener 'through an already existing terrain' (Ibid, in Zareva, 2013, p.75). At the more assertive end is the 'I' who gives their personal opinion on information discovered in the research process and who makes and is accountable for knowledge claims. However, as Yang (2014b) makes clear, we must be careful not to generalise generic rules about stance across disciplines as a speaker's techniques for conveying authorial voice may differ between the hard and soft sciences.

I would now like to discuss some practical aspects of the assessment process: namely, moderation, rater training and rating scales. After this I will analyse some studies of rater strategies and rater orientations which also used the verbal report method.

2.4. Rater Moderation

In Fulcher's (2003) overview of reliability studies of speaking assessment, he concluded that potential reliability problems could be countervailed by two things: rater training and moderating a rater's score with a second rater, 'in order to avoid the possible impact that a single rater may have on the test score' (p.141). The EAP Unit improve the reliability of their assessment by having two raters assess every ACP performance. After every performance the raters discuss their individual scores and negotiate a moderated score. The ACP is also video-recorded so that, if there is a large disparity between the raters' opinions, the performance can be third-marked. Although the focus of my study is on individual rater perceptions, the fact that the assessment is co-rated seems to be a fundamental aspect of the EAP Unit's assessment process, and this is something that I must bear in mind when I am researching rater orientations. My study will analyse raters perceiving performances *individually*, but many of these raters may be accustomed to benefitting from a second opinion on every performance.

2.5. Rater Training

Rater training is another crucial aspect of the rating process which will have a huge impact on the type of literature I will discuss later and the research questions I will pose. According to Weir (2005), rater training is ‘a systematic process to train raters to apply the rating scale in a consistent way’ (p.190). The first reason for this is to maintain internal and external rater consistency (Fulcher, 2003). Internal consistency is ‘the extent to which the same rater awards the same score to the same individual over a period of time’ (Ibid, p.139) or intra-rater reliability. External consistency, or inter-rater reliability, is the ‘extent to which two or more raters are capable of agreeing with each other on the score they award to the same individual’ (Ibid). The second reason for implementing training is to socialise the raters into a ‘common understanding of the scale descriptors’, so that they become ‘adepts’ at using the scoring system, and begin to ‘see speaking in terms of the scale they are using’ (Ibid, p.143). The phrase rater training is often used interchangeably with ‘standardisation’ as both convey the idea of bringing raters ‘into line’ (Weir, 2005).

2.6. Rating Scales

Raters of writing and speaking usually assign a score by matching features of performance to verbal descriptions which ‘describe briefly what the typical learner at each level can do’ (Upshur and Turner, 195, p.4). These verbal descriptors or rating scales can perform various functions. As well as being used to help raters make consistent decisions (rater-oriented scales) (Luoma, 2004), rating scales can also help test developers ‘select tasks for inclusion in the test’ (constructor-oriented scales) (Fulcher, 2003, p.89). They may also serve diagnostic purposes, allowing test-takers to self-assess their level and identify their strengths and weaknesses (user-oriented scales) (Luoma, 2004). The EAP Unit’s rating scale is assessor- and user- oriented. A copy of the rating scales which has the individual ‘cells’ or descriptors highlighted according to students’ scores is handed back to the students for feedback after the ACP assessment.

There are two main types of rating scale: holistic and analytic. In holistic rating scales, the rater gives a single score, or band level, which is based on ‘an overall impression of an examinee’s ability’ (Luoma, 2004, p.60 - 61). This makes it quick to read and score from the assessor’s point of view (Ibid). It is also flexible in that it allows different combinations of strengths and weaknesses to be included in one band level (Ibid). On the other hand, holistic scales are not so useful for diagnostic purposes (Ibid). An analytic scale, in contrast, ‘separates and weights different features of the test taker’s performance’ into criteria (Richards and Schmitt, 2010, p.25). Under each criteria are usually descriptors which describe performance features of that criteria for each band level. Analytic scales result in better diagnostic ability (Luoma, 2004; Knoch, 2009) and give more detailed guidance for the rater (Knoch, 2009). However, this can be a two-edged sword as it can also produce a ‘halo effect’ in that scores from one criteria can contaminate scores from another criteria, influencing the rater to give the same score across criteria boundaries (Fulcher, 2003). To counteract this, raters need robust training to make them aware of the importance of rating each aspect of the performance separately (Knoch, 2009). Analytic scales are also time-consuming to apply and their emphasis on dividing the spoken performance into a range of criteria ‘may divert from the overall effect of a performance’ (Weir, 2005, p.191). The descriptors used by the EAP Unit are analytic.

2.7. Rating Scale Design

There are two main approaches to rating scale design: intuitive and empirical. The intuitive approach, the more common of the two (Fulcher, 1996), and the type which the EAP Unit uses, involves an expert or committee of experts designing a new rating scale based on existing rating scales and/or what they 'think might be common features' at various levels of performance (Knoch, 2009, p.276). Descriptors are then drafted at different levels and the scale goes through several iterations during which other experts and pilot test-takers give their input until 'a usable formulation of the scale' is arrived at (Luoma, 2004, p.83). This is called an 'a priori' or 'thin description' approach to test design because it is built on theory rather than data (Fulcher, 2003). Intuitively-based rating scales may seem attractive to end-users, but they can also be built on a faulty or artificial construct of communication (Fulcher, 1996). Fulcher (Ibid) argues that intuitive rating scales which measure fluency, for instance, often describe a range of performance that stretches from two unrealistic extremes: from zero fluency at one end to native-like fluency at the other. This does not give a satisfactorily nuanced description of a learner's progression through different levels of fluency ability, nor is native-level fluency a useful criterion for students to aim for (Ibid).

Empirical approaches, in contrast, are based on 'thick' data of test-takers' actual language use (Ibid). An example of this approach would be Fulcher's (Ibid) development of a fluency-based rating scale. To do this he analysed twenty one transcripts of students taking the ELTS test (now IELTS) (Ibid). He coded each script for fluency features and counted these features (Ibid). Discriminant analysis was then used to find out to what extent the frequency of fluency features 'could predict the band/level into which each learner had been placed by an ELTS test' (p.213) (Ibid). Fulcher used this information to inform the design of a data-driven rating scale and found the new scale to be highly reliable and valid: that is, it gave consistent results and helped the test developer test what they were intending to test (Richards and Schmidt, 2010). This accords with Knoch's (2009) findings in the field of writing. Comparing a standard proficiency scale and an empirically-derived scale, she found that the empirically-derived scale was more explicit, helped raters arrive at a score more precisely, and reduced the likelihood of raters resorting to the 'halo effect' (Ibid). However, against these gains in terms of validity, the practical element must also be considered. As

Knoch stresses, the empirical approach took longer to develop and needed more funds and manpower (Ibid).

2.8. Issues in Rating Scale Development

Two key concerns that will occupy the developer of an analytic rating scale are how many levels to use, and how many criteria. Luoma (2004) explains that the more levels a rating scale has, the more specific feedback a rater can give and the easier it is for a student to measure their progress. However, if a rating scale has fewer levels, this makes it easier for a rater to distinguish between the levels consistently, and the rater's decision will become more reliable (Ibid). As a compromise, Luoma recommends four to six levels. For criteria, Luoma states that four to five criteria 'begin to cause a cognitive load for raters' (Ibid, p.80) and seven should be the upper limit. The EAP Unit's descriptors have six levels, but twenty seven criteria! Therefore, one of my research questions for the verbal report and the follow-up questions will be: is the rating scale manageable for the raters?

Knoch (2009) makes a pertinent criticism of many writing assessment descriptors which could equally be applied to speaking assessment: that is, they use 'relativistic wording between levels' rather than 'precise and detailed descriptions of the nature of performance at each level' (p.277). As Luoma (2004) states, descriptors need to include concrete descriptions of performance or examples of target tasks which can be performed at each level. Furthermore, criteria should be conceptually independent and the rater should not have to read the adjacent descriptor or cell (one band higher or lower) to understand what a specific descriptor means (Ibid). One way to improve the reliability of the rating scales is to avoid using qualifiers such as 'many', 'most', 'a few'' and adjectives such as 'limited', 'good', and 'fluent' (Ibid; Fulcher, 2003) which can be interpreted subjectively by the rater. This will be another aim of my research: to investigate how clear the language of the EAP Unit's descriptors is.

2.9. Studies of Rater Strategies

Very few studies have been conducted on rater strategies, and the studies that exist come mainly from the field of written assessment. Cumming, Kantor and Powers' (2002) research of rater orientations in TOEFL essays found that their raters displayed, 'complex interactive episodes of information-gathering, reasoning, and hypothesizing' before they evaluated the performance (p.73). They also found that raters used 'diverse self-control strategies' to monitor their decision-making (Ibid, p.88). Some of the codes I will use later derive from this study.

In Pollitt and Murray's (1996) research, which used the Cambridge Proficiency speaking exam as stimulus, the authors found that the raters could be divided up into two broad types: synthetic and analytic. Synthetic raters tended to use a 'preconceived….preconstructed understanding of language learners' to construct a 'holistic image of the speaker' from a few initial impressions (Ibid, p.86). Pollitt and Murray argue that these raters focus on one aspect of performance as a 'primary indicator of level' and compare the performance to a mental image of a previous test taker that they have rated at a similar level (Ibid). If the image fits the present circumstances, all traits of that 'mental image' are applied to the present performance. Analytic raters, on the other hand, were more objective, more guided by observable behaviours, and thought within, 'a strictly assessment-oriented framework' (Ibid, p.87). A similar finding is reported by Meiron (1998, in Brown, Iwashita and McNamara, 2005) who divides the two approaches into 'quasi-analytic' in which 'discrete features in the speech sample were differentially weighted to arrive at a final score', and 'Global/holistic', where 'the rater did not focus on any one specific feature' (p.6). I would therefore like to find out whether any analytic or holistic rating tendencies can be found in my data, or whether the raters are using any other identifiable strategies to cope with this task.

2.10. Studies of Rater Orientations

Fulcher (2003) points out that studies which investigate the construct validity of speaking tests, particularly verbal report studies, tend to use untrained raters. The reason for this is that untrained raters are like ‘blank slates’ who have not been contaminated by their training, and so can give useful data on the construct of tests. Trained raters, in contrast, have been socialised into the rating scale, a scale which has already been presumed to be valid, and so they ‘think in terms of the system they are using’ (Ibid, p.146).

This complex inter-relationship led me to focus my reading on studies which used trained raters as participants, because my participants in the EAP Unit were all highly trained and adept at using their rating scale. It also constrained the kinds of research questions I could ask. As Fulcher recommends, when using the verbal reporting technique with trained raters, the researcher’s focus should be, ‘to discover if the raters are applying the criteria in which they have been trained, or whether they are bringing personal construct-irrelevant criteria to the rating process’ (Ibid, p.147).

Several studies using the verbal report methodology have indeed found trained raters heeding aspects of performance that are not described in the rating scale. Orr (2002), in a study of the FCE paired speaking tests, found that accredited raters heeded many non-criterion aspects and that the rating scale was ‘heeded in different permutations by different raters’ (p.153), particularly the more complex criteria. Brown (2000) studying the IELTS interview, found that accredited raters made several non-criterion judgements, such as inferences about the examinee’s personality, test-wiseness, maturity, world knowledge and choice of topic, amongst others. May (2006) found that the raters made construct-irrelevant comments about content, such as quality, complexity, and relevance of ideas and she concluded that these features needed to be included in the rating scale. Several of the researchers concluded that their rating scales were unclear (Orr, 2002; Brown, 2000) and that their raters needed to be trained more assiduously in what the criteria did and did not consist of (Orr, 2002; May, 2006).

Another factor which might influence what raters attend to in a performance is proficiency level. In Pollitt and Murray's (1996) study of an oral proficiency test, oral proficiency was characterised in different terms at different band levels. For lower proficiency students, the raters focused on vocabulary, grammar, comprehension and pronunciation. We may infer from this that a possible reason was that the rater had to make more effort to comprehend the students' production at the lower proficiency levels and so found these features more salient (Ibid). At higher levels, raters attended more to stylistic devices and sociolinguistic competence. A correspondence can also be drawn to Brown, Iwashita and McNamara's (2005) verbal report study of the TOEFL speaking test. They found that raters made more comments about fluency and pronunciation at the lower proficiency end of the spectrum. However, linguistic resource and content were the first and second most frequent type of comment across all the band levels, and so it does not exactly mirror Pollitt and Murray's findings. Also, it must be mentioned that these two studies focused on untrained raters rather than trained raters. Nevertheless, with these studies in mind, I wanted to investigate whether a medium-level (B grade) performance from a student on an interim pre-sessional course would cause the raters to heed accuracy and fluency features as much as genre and/or criticality features, or whether one of these was predominant.

In summary, my four research questions are:

- To what extent do the raters attend to criterion features of performance?
- Do the raters focus on language features, such as fluency and accuracy, or are genre and/or criticality equally or more heeded?
- Are there any issues in how the raters use the rating scale, such as manageability and clarity?
- Can any patterns in rater strategy be construed across the data?

3. Methodology

In this section, I would like to outline two types of verbal report, briefly explain the assumptions the methodology rests on, as well as draw attention to some of its limitations. With these limitations in mind, I will explain how I conducted the research and analysed the data. Finally, I will describe the interview process which I used as a way to support and enrich the verbal report data (Dörnyei, 2007).

3.1. The Verbal Report: Definitions, Strengths and Weaknesses

The verbal report is a type of introspective study which aims to elicit data about a participant's 'thought processes involved in carrying out a task or activity' (Gass and Mackey, 2000, p.1). It is particularly useful in investigating what kind of information is heeded by a participant and what kinds of strategies they use when performing a task (Green, 1998). However, it is important to remember that these 'think-aloud' reports do not reflect the thought processes themselves, but represent 'a subset of the information currently in short-term memory' (McKay, 2009, p.222). The researcher therefore has to make an inference from the elicited verbal data to the participant's underlying cognitive processes (Ibid). The methodology rests on three assumptions: that it is possible to observe internal processes as we might observe external phenomena; that participants have access to their internal thought processes, and that they can verbalise these processes (Gass and Mackey, 2000). According to Gass and Mackey, there are two main types of verbal report: concurrent and retrospective elicitation. In concurrent elicitation, the participant introspects while enacting a task for the first time (Ibid). This is more difficult for the participant and requires extensive training and a model for them to follow (Ibid). Retrospective elicitation happens subsequent to the original task (Brown and Rodgers, 2002). It uses a stimulus such as a video or audio recording to spark the participant's memory and elicit the decisions the participant made when they enacted the task for the first time. This is the type of elicitation I will use in my study.

There are two main objections to the verbal report methodology. The first is the veridicality of the method. Several researchers have argued that major areas of cognitive processing may be inaccessible to reporting (Lumley and Brown, 2005) or that direct access to cognitive processes is impossible as these processes are mediated and warped by filters such as memory (Gass and Mackey, 2000) or a natural, human ‘tendency to make sense of whatever it is we pay attention to’ (Fulcher, 2003, p.222). From the researcher’s point of view, the data elicited from this technique is also highly inferential (Gass and Mackey, 2000) and therefore prone to idiosyncratic interpretations (McKay, 2009). This last point can be mitigated by employing a co-coder to independently code a subset of the data, so that the interpretation of the results are not biased too much by the perceptions of one researcher (Green, 1998). Intra-coder reliability can also be boosted by the same coder coding the data twice, although there is also a danger that they might make the same error twice or be biased by their memory of the first coding (Ibid). Another way to mitigate possible sources of unreliability in the method is to combine verbal reports with another research method to triangulate the data (Dörnyei, 2007; McKay, 2009; Lumley and Brown, 2005).

The second main objection is reactivity. This refers to the possibility that the participant will report what they think the researcher will want to hear (Ibid), or that the researcher’s instructions, training and examples may ‘lead’ the participant to make certain comments which are favourable to the researcher (Lumley and Brown, 2005). This is a particular problem for concurrent elicitations. For retrospective elicitations, Gass and Mackey (2000) recommend giving the minimum training necessary, so as to avoid ‘influencing or affecting the subsequently recalled data’ (p.52).

Although veridicality and reactivity are huge challenges for the verbal report researcher, they do not invalidate the method, as long as the researcher approaches the data collection and interpretation with caution (Gass and Mackey, 2000). In fact, in the field of L2 assessment, this methodology has been extremely fruitful, improving rater training and rating scale design (Fulcher, 2003).

3.2. Participants

The participants are all tutors working for the EAP Unit. They are experienced EAP tutors with, on average, fifteen years EAP experience and an average of four years experience rating ACPs. Four of the raters (Sarah, Rebecca, Rachel and Leah[1]) are Pre-sessional 3 tutors. Ruth occupied a more senior position, but often helped out in assessment rating. They all received updated rater training in May and June 2016 as a compulsory part of teaching and/or assessing the course.

These tutors were recruited through an email sent to all the Pre-sessional 3 tutors (Appendix B). Four volunteered and I also asked Ruth if she would not mind participating because of her experience of Pre-sessional 3 assessment. I asked these participants for their consent to record, transcribe and store their verbal reports and transcripts in line with the British Educational Research Association's (BERA) 'Revised Ethical Guidelines for Educational Research' (2011) (Appendix C).

3.3. The Stimulus

In accordance with the BERA guidelines, I asked consent from Pre-sessional 3 students to use videos of their ACP performance (Appendix D). Thirty students gave me consent for this. From this sample, I eliminated the high and low scoring performances. Then, in order to recreate rating conditions which were as authentic as possible, I eliminated those students who had been rated or taught by my volunteer raters. From the remaining sample, I selected the performance of Grace[2], a Chinese Msc Management student, as the stimulus. I found her performance particularly interesting because she used a case study format, a genre

[1] Pseudonyms

[2] A pseudonym

commonly assessed in business and management programmes (Nesi and Gardner, 2012). In her case study she compared the corporate culture of *Starbucks* and *Apple.*

3.4. Piloting

The piloting phase is particularly crucial to a verbal report study (Gass and Mackey, 2000; Brown and Rodgers, 2002; Dörnyei, 2007). I carried this out with three classmates, each retrospecting on my chosen stimulus with a simplified rating scale adapted from Rice, Baysen and Stetler (2004) (Appendix E). This gave me a good grasp of the timing of the different steps in the procedure and allowed me to draft and re-draft a set of instructions. These instructions would help me standardise the procedure, so that each rater received the same input from the researcher. To minimise reactivity in the verbal report, I also tried to avoid using language in the instructions which might cause the rater to focus on one feature of performance to the exclusion of others (Gass and Mackey, 2000). The final draft of the instructions is in Appendix F.

3.5. Procedure

The procedure is based on the verbal report procedure of Brown, Iwashita and McNamara (2005), Brown (2005) and May (2011). This procedure features four stages (Table 1). In the first stage, the rater watches a video of the performance all the way through without stopping and then assigns scores using the band descriptors. Next, the rater retrospects their reasons for giving the scores (Brown, 2000). This is called the 'summary turn' (Ibid). In the final stage, the rater watches the performance again, but pauses at salient points or 'wherever she/he feels some comment was in order' (Ibid, p.58). These are 'review turns': the focal point of the data. The whole session would take about an hour. I felt that any longer would have imposed on my participants' time and would have fatigued them (Green, 1998). The participant watched the stimulus on a laptop and I sat parallel to her. My main job was to prompt the participant to 'remember to pause' or to clarify their comments, but not to intervene (Green, 1998). I also took field notes of the verbal report to disambiguate any unclear parts of the transcript when I was analysing the data (Appendix N).

Stage	Rater Actions	Timing (Approx.)
1	Rater watches the performance without stopping	16 minutes
2	Rater scores the performance using the descriptors	2 minutes
3	Rater gives their scores and justifies these scores in an oral summary (the summary turn).	3 – 5 minutes
4	Rater watches the performance a second time and pauses at salient features of performance which influenced their scores (the review turns).	30 - 35 minutes

Table 1: the procedure for the verbal report

3.6. Training

I decided to implement the training through a seven minute *Youtube* video which would succinctly demonstrate the procedure. I was worried that training would take up vital face-to-face reporting time and so the video seemed a practical solution to this. As there were no *Youtube* videos available which would serve this function, I made the video myself by recording myself retrospecting on an IELTS Speaking Exam Part 2 sample video, which is available on *Youtube* (Polushkina, 2013). I chose an IELTS task because it was sufficiently different to the ACP, and therefore would not contaminate the participant's retrospection. This was then sent out as a link to all the participants. The video is available at https://www.youtube.com/watch?v=TIn5vB_uhq8.

3.7. Transcription

The summary turns and the review turns were both transcribed using a reasonably 'fine-grained transcription' (Bowles, 2010) which included pauses, hesitation markers and false starts (Appendices I and K). This was because I did not want to transcribe with a specific agenda (Richards, 2003). However, when I present the information I will tidy up the transcription to make reading easier (Ibid).

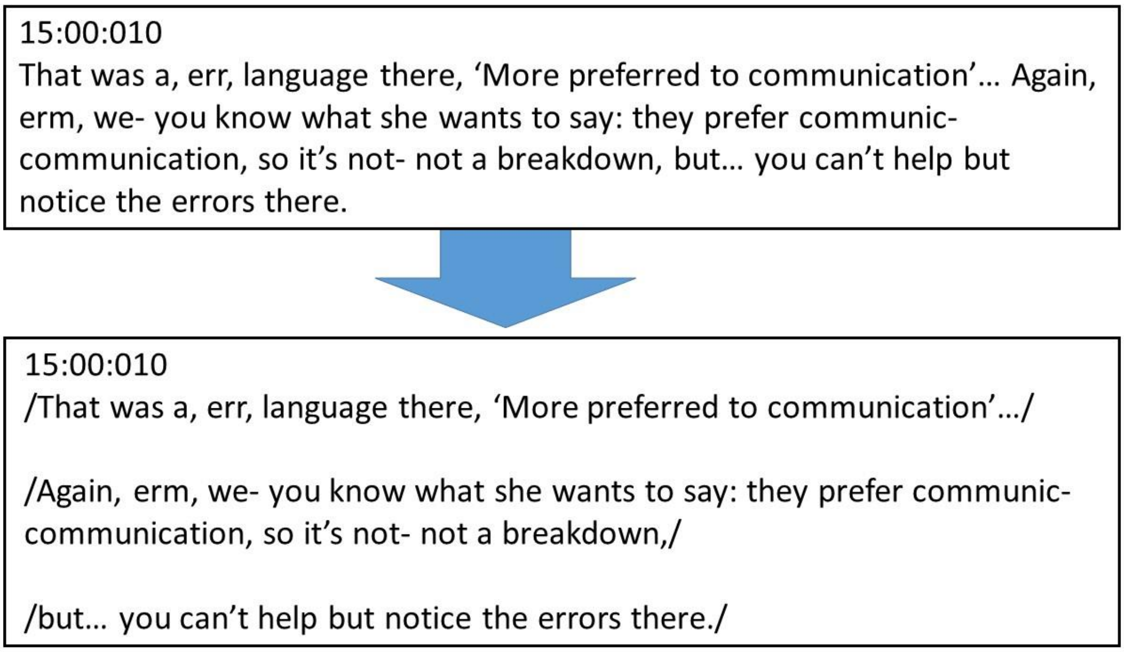
3.8. Coding

I looked through a number of studies on rater orientations in both speaking and writing assessment, but found very few code schemes[3] which I could adapt to make my own code template. The code template is a kind of prototype code scheme from which the researcher can develop a more refined scheme which fits their data more accurately (Dörnyei, 2007). However, Cumming, Kantor and Powers' (2001, in Gebril and Plakans, 2014)[4] study gave me three valuable codes which would form a basis of my code template: interpretation, self-monitoring and judgement processes. I interpreted judgement processes as the evaluations of performance made by the raters. I adapted self-monitoring into a broader code: reflection and self-monitoring. I interpreted this as a type of metacognitive process in which raters commented on or guided their own rating behaviour. I adapted interpretation processes into 'processing oral and visual input'. This would describe how the rater verbalises the process of decoding Grace's language, structure and content. The other codes emerged organically from the coding process.

The summary turn and review turn transcripts were then segmented into 'thought units' (Brown and Rodgers, 2002). Each thought unit represented a single, definable rating process and could be a clause, sentence or a whole turn (Green, 1998) (Figure 5):

[3] May (2011) has published a coding scheme for a verbal report of a paired speaking task, but this is very much student-centred, rather than rater-centred, and does not fit the ACP task well.

[4] This study was carried out on raters of academic writing. In general, there seem to be more rater orientation studies in the field of writing than in speaking.



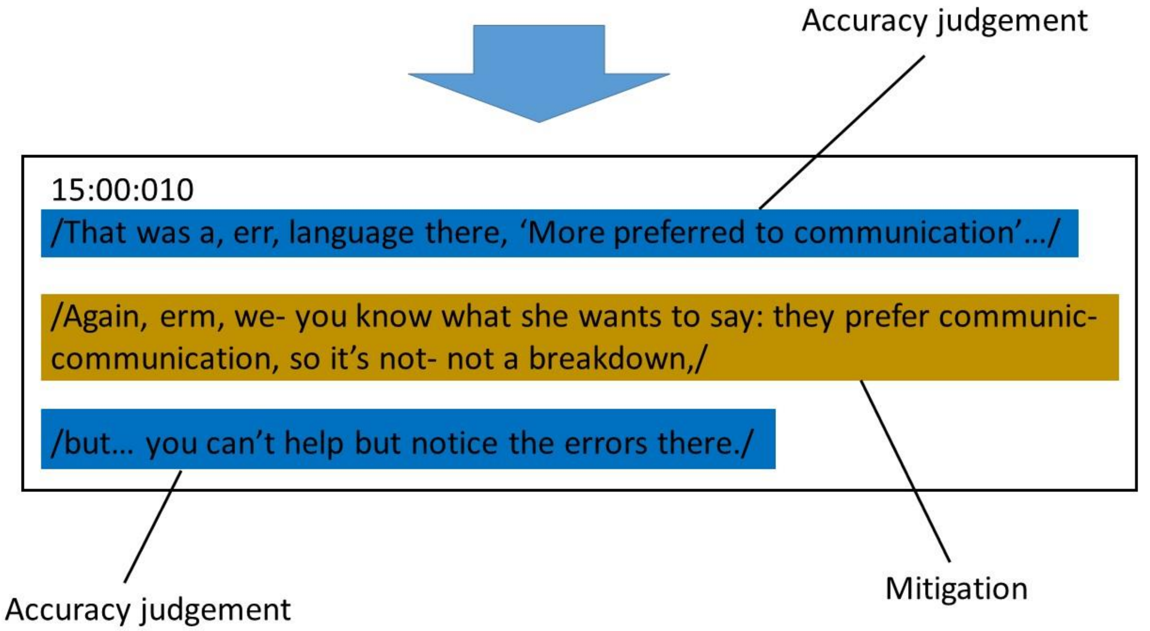
15:00:010
That was a, err, language there, 'More preferred to communication'… Again, erm, we- you know what she wants to say: they prefer communic- communication, so it's not- not a breakdown, but… you can't help but notice the errors there.

15:00:010
/That was a, err, language there, 'More preferred to communication'…/

/Again, erm, we- you know what she wants to say: they prefer communic- communication, so it's not- not a breakdown,/

/but… you can't help but notice the errors there./

Figure 5: the segmentation process



15:00:010
/That was a, err, language there, 'More preferred to communication'…/

/Again, erm, we- you know what she wants to say: they prefer communic- communication, so it's not- not a breakdown,/

/but… you can't help but notice the errors there./

Figure 6: the coding process

The codes in the coding template were applied to these thought units (Figure 6). More codes emerged as I coded the first two transcripts and others were discarded. In this way, my code scheme developed iteratively (Dörnyei, 2007; Bowles, 2010). A particularly important unit of my analysis was the judgement process. All judgement thought units were matched to the criteria in the rating scale to analyse the construct relevance (Fulcher, 2003) of the raters' evaluations. These were counted for each rater. Other processes such as reflection and self-monitoring were also quantified so that patterns could be construed across the transcript and rater strategies inferred from these patterns (Green, 1998). The quantitative results are presented in Appendix L.

I wanted to create a comprehensive code scheme which would describe every thought unit in my data. Having a broad range of codes has the advantage that it allows the researcher to make specific inferences. However, a wider range of codes also makes the data more susceptible to idiosyncratic coding (Green, 1998). Because of this, I could see that I would need to get another coder to look over a part of the data, to make sure I was not coding too subjectively. I had a classmate help me with this and we spent about an hour independently coding two pages of review turn transcript. We found that we agreed on 58% of the codes that I had assigned these two pages, which is a low reliability score. However, I felt the process was flawed for many reasons. I would have liked to code more of the data – Bowles (2010) recommends 10 to 25% - but I only had an hour with my classmate. There was also very little time to explain the study to her so that she had a full understanding of the codes and the aims of the codes. Explaining the rating scale was impossible too, as this would have taken even longer. Consequently, I mainly relied on the follow-up interviews and re-coding the data myself to boost the reliability of my study. The final coding schemes for the summary turn and the review turn are in Appendices H and J.

3.9. Follow-up Interviews

The follow-up interviews happened two to three weeks after the verbal reports. I noticed during the verbal reports that raters would only give me their opinions about the rating scale when they were 'off record', usually after the retrospection. They seemed unwilling to criticise the descriptors, and by implication, the EAP Unit. I therefore decided to hold the interviews in the university café, and not to record the conversation, but instead take notes as close to verbatim as possible. The aim was to encourage the participants to speak freely. The interviews were structured in that I had an interview schedule for each rater and targeted specific topic areas ‘in a well-defined domain’ (Dörnyei, 1997, p.135), i.e. rater strategies, opinions of the rating scale and the presence of a co-marker. I also wanted to include a question which would enable the participant to reflect on the success or difficulties of the think-aloud procedure, as this would give me good feedback on the reliability of the procedure (Lumley and Brown, 2005). I kept strictly to a time limit of half an hour because I felt I had already imposed on their time. The interview schedules had a common core but the questions were sometimes phrased differently according to the participant. This was to follow up on aspects of interest that emerged from the verbal report. A sample interview is in Appendix M.

3.10. Key to results

In the next chapter, I will mainly use data from the verbal report, but I will also weave in data from the interviews and field notes. To make it clear where the data was derived from, I will use the following code:

RT = Review Turn

ST = Summary Turn

INT = Interview

FN = Field Notes from verbal report (including annotated rating scale)

4. Results and Discussion

This chapter reports on the results from the verbal report procedure. Each of the research questions will be dealt with in turn; causes of the results will be inferred and implications for assessment practice will be suggested. The first two sections mainly draw on data elicited from the verbal report. In sections three and four, information from the interviews is weaved in more frequently.

4.1. Research Question 1: To what extent do the raters attend to criterion features of performance?

As mentioned in the previous chapter, the thought units involving a judgement, either positive or negative, were counted and compared to criteria on the rating scale. Table 2 shows that raters overwhelmingly heeded construct-relevant (Fulcher, 2003) information in their review turns, applying criteria from the rating scale in 93% of their total judgements. This is in stark contrast to the studies of Brown (2000), Orr (2002) and May (2006), where the trained raters often deviated from the rating scale, frequently applying their own frame of reference to their evaluations. The results in this study may have been due to the significant amount of rater training undertaken by these participants (4 times on average for the ACP assessment), making the raters thoroughly socialised in using the rating scale (Fulcher, 2003). We might also attribute this to the fact that the twenty seven criteria in the EAP Unit's rating scale are comprehensive, and leave little room for raters to attend to non-criterion features.

	Sarah	Rebecca	Rachel[5]	Leah	Ruth
Total thought units for judgement processes	60	45	25	45	36
Criterion judgements	54	43	25	41	32
Non-criterion judgements	6	2	0	4	4

Table 2: total thought units in the review turns which express a judgement

4.2. Research Question 2: Do the raters focus on language features, such as fluency and accuracy, or are genre and/or criticality equally or more heeded?

Research from Pollitt and Murray (1996) and Brown, Iwashita and McNamara (2005) suggested that raters may attend to language features as much as, or more than, genre and criticality features in a medium-scoring performance, particularly from a student in the middle stages of her pre-sessional development, such as Grace. However, this was not the case. As Table 3 shows, genre and criticality criteria were heeded more than language features, although with the qualification that pronunciation, grammatical accuracy and fluency were particularly salient to Ruth, Rebecca and Leah respectively. The criticality criterion 'Depth of Analysis' was particularly salient for four of the raters, ranking as either the most frequent judgement or second most frequent. 'Use of sources' was also frequently heeded, ranking among the five most frequent judgements for Sarah, Rebecca and Leah. The fact that the raters orientated to these two criticality criteria so strongly speaks of the importance that critical thinking holds in the EAP domain (De Chazal, 2014). This is reinforced by Sarah who praises the fact that the EAP Unit's descriptors give equal weight to criticality as to language **[INT]**. As a non-native speaker herself, she believes that this sends a positive message to pre-sessional students that language is not the sole or primary determiner of score **[INT]**.

[5] Rachel's review turn elicitation was shorter because she digressed halfway through the retrospection to talk about points of interest raised by the performance and the rating process in general.

Sarah	Rebecca	Rachel	Leah	Ruth
Depth of analysis (15)	Depth of analysis (6)	Visual support (5)	Audience engagement (7)	Depth of analysis (5)
Visual support (7)	Signalling (5)	Depth of analysis (5)	Ability to explain (7)	Audience engagement (4)
Audience engagement (6)	Use of sources (5)	Ability to explain (4)	Fluency (4)	Pronunciation: word level (3)
Ability to explain (5)	Structure (4)	Audience engagement (3)	Depth of analysis (4)	Ability to explain (3)
Use of sources (4)	Grammatical accuracy (spoken) (4)	Structure (2)	Use of sources (4)	Grammatical accuracy (spoken) (2)

Table 3: the five most frequently evaluated criteria according to each rater. Number of thought units in brackets. Genre criteria in red, criticality in blue, language in yellow.

In the next two sections, I would like to analyse some commonly heeded criteria from the criticality and genre categories of the rating scale and set the raters' comments within the theoretical context described in the literature review.

4.2.1. Criticality Judgements

The most frequently heeded criterion was 'Depth of analysis'. This was often evaluated by raters when Grace simply described what corporate culture was or what *Starbucks* and *Apple* did in terms of corporate culture. This could be likened to the first level of Bloom's taxonomy of cognitive development: she is recycling knowledge but not deconstructing that knowledge or combining it with other sources (De Chazal, 2014):

> **[RT] Rachel**: 'Could have done more […] there's all sorts of….detail, that would have substantiated that but just saying- it's just you know- it's part of what I said about- there's a first line of analysis and then nothing, nothing more'.

The ability to evaluate sources was particularly salient in many of the verbal reports. Grace used images of smiling *Starbucks* employees and a promotional video to support the claim that *Starbucks* has an open communication system in its corporate structure and that this creates happy staff. Several of the raters pointed out how she failed to realise the hidden agenda of these sources:

> **[RT] Rebecca**: 'She's just making the point about happy employees, so it seems quite a simplistic one […] this is a video produced by *Starbucks* to promote their product, so what does it actually show? Not really an academic source'.

Evaluations about 'Depth of Analysis' were often interspersed with evaluations about 'Use of Sources' and the retrospections showed how closely the two are linked in EAP. Here, the rater highlights Grace's lack of substantiation for claims, which could have been achieved with academic sources or statistics:

> **[RT] Rebecca**: 'When she's saying, "Most people say…" who are those people? You know, this is where a source would be… useful. You would want to be convinced by what she's saying'.

Raters' evaluations about depth of analysis also encompassed Grace's inability to see both sides of an argument and to take an objective view of the arguments surrounding big business and corporate culture. At times, I found this difficult to distinguish from the criticality criterion, 'Attempts to assess strengths, weaknesses, limitations and potential solutions', but the raters did not seem so much to be looking for an evaluation of real-world business problems: rather they seemed to be looking for Grace to weigh evidence from different sources:

> **[RT] Ruth**: 'That would have given her an opportunity to introduce the counter-argument I mentioned before… you know, if this is what Apple say about themselves, is that really true? Isn't this just one side of the argument?'

Given the primacy of stance to criticality (De Chazal, 2014; Cotton, 2010), I was surprised that there were not more judgements relating to the 'personal stance' criterion, but this may have been due to a lack of personal voice in the overall performance. For example, Sarah praises Grace's use of the assertive self-mention (Zareva, 2013):

> **[RT] Sarah**: ‘OK, this is good, you see, she said, “I think”. For me, as an examiner, that’s a signal she’s providing her stance, her position, and that’s a good thing’

This shows Grace projecting the identity of the ‘opinion-holder’ which is valued in Western academia (Ibid). A skilful manipulation of authorial voice is crucial in academic discourse as it allows the presenter to ‘stamp their personal authority onto their arguments or step back and disguise their involvement’ (Hyland, 2006, p.176). However, an assertive opinion needs to be developed and substantiated. For example, Sarah revises her earlier opinion a few turns later when the student fails to develop her line of thought sufficiently:

> **[RT]** ‘There is no real explanation coming, she’s kind of reiterating what she’s just discussed’.

Rebecca goes through the same change of heart:

> **[RT]** ‘Ah, interesting opinion there… from her […] well again, why does she think that? What does it show?’

4.2.2. Genre Judgements

As can be seen from Table 3, the genre criteria ‘Visual support’, ‘Audience Engagement’ and ‘Ability to explain’ were all in the five most frequent judgements according to four of the raters. ‘Ability to Explain’ was particularly salient, representing 22 thought units across the transcripts. However, this criterion often seemed to blur into the more general critique about a lack of substantiation for claims and a lack of analysis:

> **[RT] Sarah**: ‘Like, is it a definition? She’s all about it’s very successful, now we don’t know….why: any statistics, anything that shows that, let’s say, this company really is truly making lots of money and it’s like one of the top globally, so no evidence here’.

> **[RT] Rebecca**: ‘She’s basically said these are the tools for creating business success in different ways, without actually saying how that happens’.

Consequently, I saw these judgements as inextricably linked to the general critique about a lack of analysis. In hindsight, it would have been interesting to follow this up in the interviews: did the raters see this criterion as more part of the criticality category or the genre category? Unfortunately, my time in the interviews was concentrated on asking questions about the rating scale and rater strategies.

The raters all showed an appreciation of the importance of tenor (Halliday and Hasan, 1989) in the ACP speech event, particularly in the set-up or contextualisation phase of the ACP (Ventola, 2002). Here, the speaker used kinesic resources such as eye contact, gesture and smiling (Morell, 2015), as well as humour, 'to resolve tensions and construe a relationship of solidarity or rapport with their audience' (Hood and Forey, 2005, p.292). This can be seen especially in the positive rater reactions to Grace's initial attention-grabbing techniques (Morell, 2015):

> **[RT] Sarah**: 'That's very funny; that's also quite nice for an international student to be able to say, "I'll stand by your evil boss", so she's sort of, telling already what that might be about: so that's good as well'

Raters were also particularly sensitive as to whether the student maintained eye contact or read from her notes or the screen (both described in the 'Audience engagement' criterion). Their evaluations of this criterion were positive in the early stages:

> **[RT] Rachel**: 'So, already, I can see that she's facing the audience, which is good. She's smiling and she engaging with them, so I'm already favourably influenced by that'.

But these judgements often soured in the later stages of the presentation as Grace struggled to maintain the flow of her presentation:

> **[RT] Leah**: 'I found this bit really weak. She's kind of reading and I'm not really sure what she's talking about'.

'Visual support' was another commonly heeded genre criterion. Grace's visuals seemed to have an illustrative function (Morell, 2015), but she did not clearly identify what the visuals represented or contextualise these images within her wider topic (Cassily and Ventola, 2002). Also, reiterating the point made earlier about maintaining an objective view, the contextualisation of the images seemed to reinforce one exclusive perspective, which was the company's perspective:

> **[RT] Sarah**: 'This visual, it is not really supporting what she is trying to say, or, is actually supporting the superficiality of this, I don't know, not analysis yet'.

4.2.3. Criticality and Genre Judgements: Summary

The raters' evaluations were dominated by criticality judgements, and many of these criticality judgements seemed to be making the same comment about a general lack of analysis in the performance, but from different angles. Genre criteria were also frequently attended to. These heeded criteria reflected authentic academic processes such as establishing rapport with an audience, using notes effectively, and contextualising visuals. These task-based criteria all reflect the performance construct (Chapelle and Brindley, 2010), the academic conference presentation, and this is something that Sarah highlights as a positive feature of the rating scale **[INT]**. However, from a coder's perspective, I found the genre criteria 'Ability to explain' and 'Visual support' difficult to distinguish from criticality judgements. If the rating scale were to be revised, I would be inclined to incorporate the 'Ability to explain' criterion into the criticality criteria, but I would need to justify this with more data from more verbal reports to see if the same process happened with other performances. I would also need to ask for raters' opinions about how they personally perceive the differences between the criticality and 'ability to explain' criteria.

4.3. Research Question 3: Are there any issues in how the raters use the rating scale, such as manageability and clarity?

In this section I would like to investigate some discrepancies in the way raters interpreted several criteria in the verbal reports and then widen my focus to look at raters' general opinions of the rating scale.

Sarah expressed doubts about how to rate the genre criterion, 'Clarifying links between the student's future discipline and the conference theme' **[INT]**. She was unsure how to rate something that was not mentioned explicitly in performance, other than Grace saying she was an MSc Management student and that she was focusing on two business case studies **[INT]**. After the summary turn, Sarah marked the student in the 'D/E' band for this **[FN]**.

Then, when she watched the performance a second time she noticed that there *was* a tenuous link to the conference theme:

> **[RT]** ‘OK, so actually she does have a word “power”: actually there was “powerful” in the title, so that’s positive. Again, perhaps, she could’ve elaborated on that: she could’ve said more about it, but anyway, that’s good’.

But she keeps the score in the ‘D/E’ band **[FN]**. Rebecca also assigns a ‘D/E’ grade. Ruth, on the other hand, is much more lenient, giving Grace a ‘B’ **[FN]**. She seems to see a link made implicitly between the presentation content and the theme:

> **[RT]** ‘It was nice how she went through the corporate structure point by point and explaining how this is linked to the success of the company and “Power”, I think she attempted to link it to the theme’.

Although this too is qualified at the end of the review turns:

> **[RT]** ‘She could link it a little bit more consistently to power, I mean she mentioned the theme in the beginning but she could refer back to that as well’.

This discrepancy could be symptomatic of the phenomenon observed by Fulcher (2003) and McNamara (1996), where raters, despite receiving training, differ in how severely they apply the criteria. This is cogent, but I also think that this discrepancy might have something to do with how the rating scale is phrased. I believe that, in this case, the criteria need to be more distinguishable from each other and describe more effectively the observable behaviour that students exhibit at different levels (Luoma, 2004). For example, is just mentioning the conference theme in the presentation title enough? Does the student need to make it clear in their introduction how the topic relates to the theme and to their discipline? Sarah’s use of the word ‘move’ suggests that it should:

> **[ST]** ‘She missed the main step or main move we encourage all students to do in Pre-sessional 3. I didn’t spot any specific attempt to point out for the audience the links between her discipline, the presentation and the theme’

This kind of ambiguity could be resolved by an empirically-derived rating scale which would describe ‘observed learner behaviour’ at the different levels, as opposed to ‘postulated or normative notions of how learners ought to behave’ (Pieneman et al, 1985, in Fulcher, 2003, p.98).

A similar type of discrepancy involved the students' use of sources. Students are required to include a page of bibliography in their presentations, but they do not have to reference the text on the slides (in-text referencing). As Sarah says, this is a source of controversy amongst the staff **[INT]**, and I felt that this ambiguity was reflected in the verbal reports. The first problem seems to be that there are two criteria describing similar features: 'Sources fully and correctly referenced' in the genre category, which relates broadly to whether the student has fulfilled academic conventions, and 'Use of Sources' in the criticality category, which describes, in a more nuanced way, what students can do at different levels in terms of use of evidence and criticality. Leah has doubts about the 'Sources correctly referenced' criterion, expressing the need for a co-marker to clarify **[ST/FN]**, but she gives a C for 'Use of sources' **[FN]**. Sarah and Rebecca perceive both criteria severely, giving Grace a D/E for both **[FN]**, despite Grace having an – albeit short - bibliography. Rebecca's summary turn suggests that she sees the 'Sources correctly referenced' criterion as solely a measure of in-text referencing and Sarah may have thought along the same lines. Ruth, on the other hand, scores 'Sources correctly referenced' as a B, and 'Use of Sources' as a C **[FN]**. She echoes that it is not obligatory to use in-text references, but recognises that it 'would have added to the value of her (Grace's) criticality' **[RT]**. Again this seems to be a situation where an empirically-driven approach would have been able to define specific behaviours at different levels more effectively. There also seems to be a need for consensus amongst the staff about what is expected in regards to in-text references. It seems unfair that students should be marked down for not doing something that is optional.

In a similar vein, Rachel mentioned how difficult it was to give a mark for the criticality criterion, 'The reporting of other people's ideas and stances' **[RT]**. If the feature does not occur in the performance, and it is not a generic feature of the ACP, how can you score the absence of it? Do you give it an 'E'? This seems too severe **[RT]**. Rachel's solution was pragmatic: she pays more attention to the most salient criteria on the rating scale and fills in others 'because you have to' **[RT]**. This coping strategy will be analysed in more depth in the next section. Ruth and Rebecca agreed that this criterion was difficult to score in this case

[INT], with Ruth adding that it is not always clear what to do if the student does not show any evidence of a criterion **[INT]**. In this case, she emphasises the role of the co-marker as a way to clarify how to score the criterion **[INT]**. This seems to be a common strategy among this group of raters, as both Sarah and Leah mentioned the co-rater's role as a kind of safety-net when they are unsure how to score or interpret a criterion **[ST/INT]**. Rebecca explains that, in cases where she finds it hard to score the feature, she looks at the sub-headings on the left of the rating scale which group the criteria into sub-categories: for example, 'The reporting of other peoples' ideas and stance' is grouped under the sub-category, 'Stance and Argument' **[INT]**. If time pressure means she cannot score all of the criteria, she makes sure that she has scored as many criteria as possible within one sub-category **[INT]**. Again, I feel this should not need to happen, and I believe that a data-driven rating scale should highlight more effectively what features are present or absent at different levels of performance.

One of the key aims of the follow-up interviews was to gauge raters' opinions about the rating scales. Given Luoma's (2004) recommendation for a maximum of seven criteria, the EAP Unit's descriptors seemed over-taxing, particularly when one considers that raters have a short time to rate each performance and are rating ACPs for the whole day. This is borne out by four of the raters who agreed that there are too many criteria and that the scale is unmanageable at times **[RT/INT]**. Three raters also stated that the criteria are too detailed **[INT]**. Rachel recommends 10 to 15 criteria as a manageable amount **[RT]** and Sarah, finding some of the Genre criteria 'a bit obvious', recommends 'boiling these down' to two or three features **[INT]**. However, seen in another light, the inclusiveness of the descriptors can also be a strength as three raters mentioned that the division between the three categories, genre, criticality and language, is clear and useful **[INT]**. The broad range of criteria also serves an important diagnostic function (Luoma, 2004; Knoch, 2009) as highlighted cells, along with a small box of feedback comments at the top of the descriptors is a valued form of feedback for the EAP Unit. This centrality of the descriptors to the feedback process was highlighted by both Sarah and Ruth in their interviews.

In terms of the language of the descriptors, the frequent use of vague qualifiers such as ‘generally’, ‘broadly’ and ‘mostly’ to distinguish between the band levels was also mentioned by four of the raters **[INT]**. As Rebecca makes clear, the relativistic nature of their interpretation (Knoch, 2009) requires another rater to clarify the meaning of several descriptors **[INT]**. The raters also found many qualifiers did not describe a nuanced gradation in level. For example, Ruth mentions that for ‘Visual support’, there seemed a big gap in denotation between ‘mostly effective’ (B) and ‘partially effective’ (C) **[INT]**. She made a quick calculation in her head and decided that more than half of the performance was successful in terms of visual support and therefore chose ‘mostly effective’ **[INT]**. Leah mentioned a similar gap between the C and D levels of the criterion ‘Pitched at an academic audience’ **[INT]**. There did not seem to be much difference between ‘Mostly pitched’ (B) and ‘Broadly pitched’ (C), but at D, there is a big drop-off to ‘May not be pitched’ **[INT]**. A data-driven rating scale should describe more clearly concrete tasks that students can perform and give examples of what students might say at different levels (Luoma, 2004), but I am not sure if a multi-componential EAP rating scale would be able to avoid qualifiers completely. It could be argued that less levels would reduce these types of ambiguities, but we established in the literature review that the EAP Unit’s rating scale was within the upper limit of levels recommended by Luoma (Ibid): that is, 6 levels. It seems, therefore, that if the EAP Unit continues with 6 levels for each criterion, they might have to rely on a co-rater to disambiguate vague qualifiers.

4.4. Research Question 4: Can any patterns in rater strategy be construed across the data?

In the first part of this section, I would like to analyse some of the decision-making patterns which were common to all the raters. To do this, I have divided these processes into macro-processes such as reflecting and self-monitoring, judging and processing oral and visual input (Cumming, Kantor and Powers, 2001), as well as other codes which I developed myself. The reflecting and self-monitoring process was a particularly useful code from which to infer strategy-use and so I subdivided this into a subset of micro-processes such as predicting, making hypotheses and revising hypotheses. This division is illustrated in Figure 7.

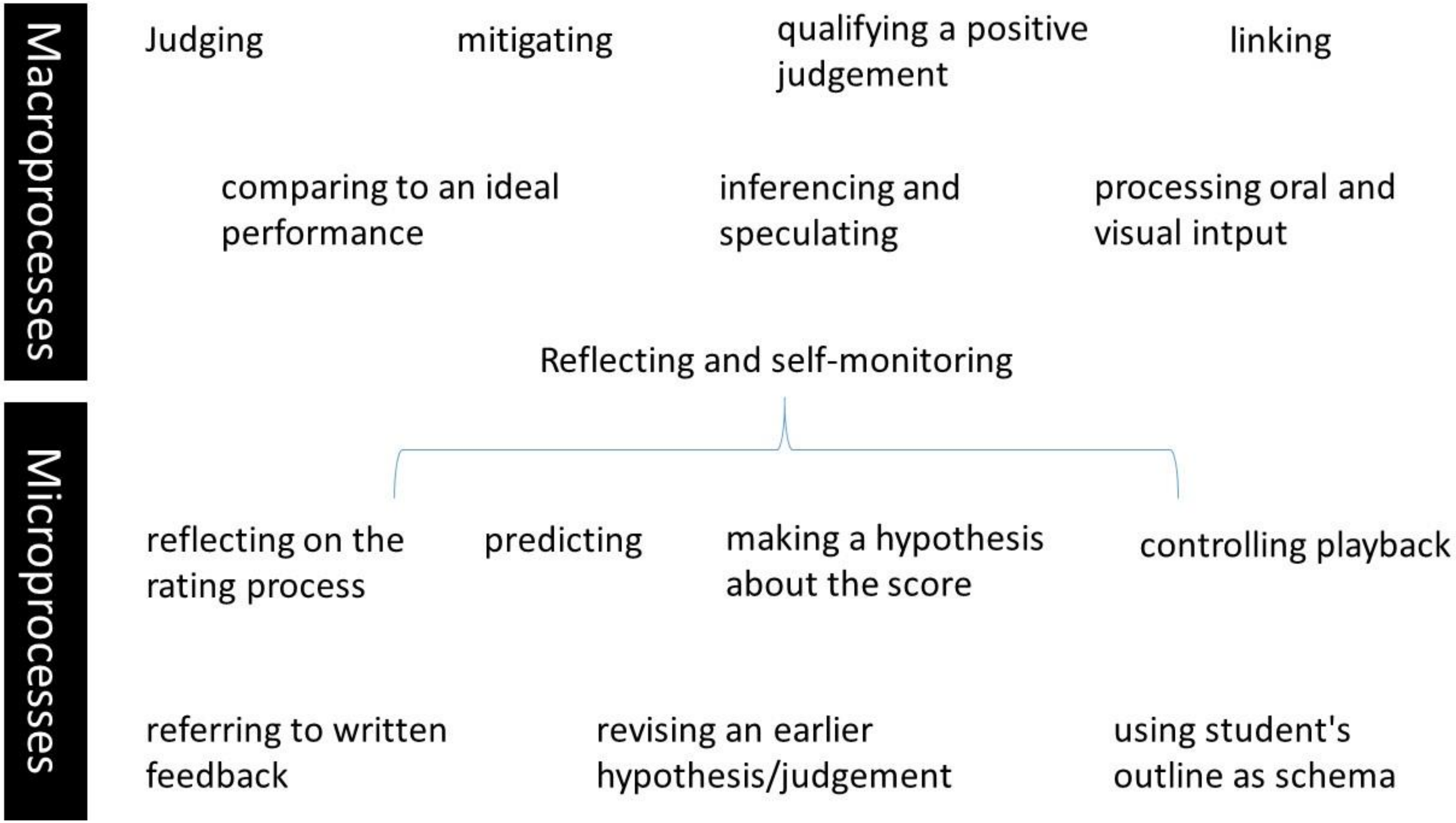

Macroprocesses
Judging
mitigating
qualifying a positive judgement
linking
comparing to an ideal performance
inferencing and speculating
processing oral and visual intput
Reflecting and self-monitoring
Microprocesses
reflecting on the rating process
predicting
making a hypothesis about the score
controlling playback
referring to written feedback
revising an earlier hypothesis/judgement
using student's outline as schema

Figure 7: macro-processes and micro-processes for inferring rater strategies

As can be seen from Table 4, the frequency of decision-making processes inferred from the verbal report varied according to rater, making it hard to discern patterns of strategy use. Moreover, Sarah and Rachel seemed particularly apt at verbalising their decision-making processes, while the other raters focused mainly on judging the performance. This can be seen in the high amount of 'reflecting and self-monitoring' thought units coded in Sarah and Rachel's transcripts. For this reason, I would like to focus on the retrospections of Sarah and Rachel in the next section.

Sarah	Rebecca	Rachel	Leah	Ruth
judgements (56)	judgements (43)	judgements (25)	judgements (41)	judgements (32)
reflecting and self-monitoring (26)	processing oral and visual input (7)	reflecting and self-monitoring (12)	processing oral and visual input (13)	comparing against ideal performance (10)
comparing against the ideal performance (10)	Mitigating negative judgements(5)	comparing against ideal performance (7)	mitigating negative judgements (7)	reflecting and self-monitoring (6)
processing oral and visual input (5)	qualifying a positive judgement (5)	processing oral and visual input (6)	reflecting and self-monitoring (6)	processing oral and visual input (6)
linking (5)	comparing against the ideal performance (4)	inferencing and speculating (5)	comparing against ideal performance (3)	qualifying positive judgement (6)
inferencing and speculating (3)	reflecting and self-monitoring (4)	linking (3)	qualifying positive judgement (3)	inferencing and speculating (5)
mitigation (2)	inferencing and speculating (3)		inferencing/speculating (1)	Mitigating negative judgements (5)
			linking (1)	Linking (1)

Table 4: Rater processes. Number of thought units in brackets.

'Comparing to an ideal performance' was a frequently coded process in three of the raters' transcripts. In this process, the rater compares a feature of the performance to a mental image of an ideal performance, or imagines how the rater would have performed in the same situation. This comparison process could be a single sentence or extend across a whole turn, weaving in and out of judgement thought units:

> **[RT] Sarah**: 'That's, of course, obvious and that's probably something we all know. Again, should she, could she have mentioned what exact goals, how to achieve those, how not to compromise?'

'Mitigation' was another process which was often inserted beside or within judgement thought units. Rebecca and Leah did this frequently by, for example, recognising the student's effort despite poor criticality:

> **[RT] Rebecca**: 'I mean, she's trying her best to explain…what these terms are, but when you're not clear why they're relevant. Yeh'.

Or by trying to understand the student's perspective:

> **[RT] Leah**: 'OK, I mean, I'm sure it- I'm sure there is a link, but I think it needed to be articulated a bit more'.

It must be said, however, that these mitigation processes do not mean Rebecca and Leah were more lenient. On the contrary, their overall scores were in line with their colleagues (C and B/C respectively).

Raters made several inferential observations, but these episodes were rare in the verbal report and I would not claim this as a sign of rater unreliability. Rather, it seemed to be a human reaction which raters were occasionally susceptible to. Interestingly, inferences were often elicited by Grace's hesitations, which were usually ascribed to a lack practice:

> **[RT] Rachel**: 'She's probably not practised that bit, so, she's lost and now she's jumping straight to the outline'.

This could well be true, but, as Leah pointed out in her summary, hesitation could also be a result of the student having memorised a script and searching in her head for her next line:

> **[ST]** 'I was worried about those patches because if the whole thing was learned and that was what she was really capable of, that was really quite weak'

A particularly salient process in Sarah's elicitation was to make links between the performance and the requirements of the next pre-sessional course or the student's future discipline. This might be due to Sarah's experience of having taught business specialism courses **[RT]**. Through these links, she drew attention to the gap that the student still has to bridge to be ready for academic performance in her subject discipline:

> **[RT]** 'But for a really good student in business, she should have focused perhaps on this comment'

'Reflecting and Self-monitoring' was a macro-process that I was particularly interested in and it was the first or second most frequent process coded in the transcripts of Sarah, Rachel and Ruth. As I mentioned before, this was subdivided into several micro-processes (Table 5):

thought units	**Self-monitoring and reflecting micro-process**
24	reflecting on the rating process, including reference to the rating scale or score
10	Predicting
9	making a hypothesis about the score
4	referring to written feedback
3	revising an earlier hypothesis/judgement
2	controlling playback
1	using student's outline as schema
1	using world knowledge

Table 5: Total self-monitoring and reflecting micro-processes across raters.

'Reflecting on the rating process' was the most common micro-process within the larger 'Reflecting and Self-monitoring' macro-process. This often took the form of the rater thinking aloud about her rationale for rating decisions:

> **[RT] Leah**: 'I'm already beginning to wonder what it's going to be about and what exactly she's going to address which is why I'm not keen on putting her in a high score here for criticality'

I will draw on this micro-process heavily when I discuss the strategies of Sarah and Rachel in the next section.

Predicting and establishing expectations of how the student's performance was going to unfold was another salient micro-process for Sarah, Rachel and Leah:

> **Rachel**: 'Now what I'm looking for now is, OK […] she's taken it to this point, and I'm now looking for an elaboration around this'.

Assigning an impressionistic or hypothetical score was another common micro-process, but this was most salient in the transcripts of Sarah and Rachel and so I shall return to this in the next section.

In summary, the verbal reports showed that these raters were not automatons: they mitigated criticisms and made unfounded judgements about features of the performance, such as hesitations, but these processes did not seem to influence the severity or the fairness of their judgements. A particularly salient process was to compare aspects of the performance to a mental image of the perfect performance, and it would be interesting to see if raters exhibit this process in other speaking assessment tasks. As in Cummings, Kantor and Powers' (2002) study, the raters exhibited 'complex interactive episodes of information-gathering, reasoning, and hypothesizing' (p.73), but two of the raters were particularly apt at verbalising these decision-making processes. I will focus on their elicitations in the next section.

4.4.1. A Comparison of Two Raters

In this section, I would like to compare the decision-making processes used by Sarah and Rachel and see if these match any of the models of rater strategy mentioned in the literature review. I will draw on data from the verbal reports and the interviews to construct a fuller picture of how each rater enacts this task. In the final section, I will try to trace patterns from Rachel and Sarah's data to the information given by the other raters in their interviews.

In Sarah's verbal report, I inferred a range of 'self-control strategies' (Ibid). In particular, she frequently made mental or hypothetical scores for criticality and language, which she challenged or confirmed as the presentation developed:

> **[RT]** 'So see, in my mind this already goes towards B in criticality, roughly, because again the presentation is not over, but I'm already not really seeing, but again, obviously, of course, the beginning, so let's give it… a try'.

And then eight minutes later:

> **[RT]** 'So criticality is actually more or less at the C in mind, but again, I'm not closing the case'

Sarah describes how she can form a hypothetical score for genre quite early in the performance, by only focusing on two or three key features, but for criticality her mental score is not finalised until the end of the performance **[FN]**. She can also develop an impressionistic score for language reasonably quickly in the performance **[FN]**. To describe her processing of language features, she uses the metaphor of the mental meter which oscillates according to the quality of the student's production:

> **[RT]** 'In my head is this… kind of balance, you know, like, "Is this like more than 5, or is like less than 5?"'

According to Sarah, this meter is particularly important for assessing language because, due to the lack of criticality in the majority of the Pre-sessional 3 students' performances, language can often become a key determiner of overall score **[INT]**.

Overall, the way Sarah approaches rating is extremely analytic, particularly in the way she mentally separates language and criticality into two 'meters' in her head (Meiron, 1998, in Brown, Iwashita and McNamara, 2005). She is also extremely 'assessment-oriented' (Pollitt and Murray, 1996) in the way she distinguishes between non-criterion and criterion judgements:

> **[RT]** 'I'm thinking she is trying hard. Again, "trying" is not really listed in the descriptors'

And in the way she strives to give a fair and consistent rating to the student:

> **[RT]** 'If you detach yourself from what's written here [pointing to the descriptors], you automatically give actually unfair advantage to students'.

Rachel shows some similarities and differences to Sarah in terms of approach. Rachel also scores the categories in her own sequence rather than the sequence they are presented in the descriptors, i.e. genre, criticality, language **[RT]**. This is a deliberate strategy to make the rating process more manageable **[INT]**. She states that she focuses on language, especially pronunciation first, as this is easiest to score and 'is likely to be unchanging throughout the presentation' **[RT]**. Criticality however is something which 'builds', and so this needs to be graded last. She emphasises, however, that these mental scores are not set in stone and can be reviewed if the performance 'suddenly becomes sparkling' **[RT]**.

Unlike Sarah, Rachel seems to favour a more holistic view of rating. Because she has seen so many ACPs, she has a clear mental schema of what a high-, medium- and low-scoring performance consists of **[INT]**. During the performance she forms a mental score and then tries to 'pin' this global impression to the descriptors **[INT]**. Like the synthetic raters in Pollitt and Murray's (1996) study she fits the criteria scores to suit her holistic mental image. However, she also uses bottom-up processes to check the details of the performance or to make revisions to her global view **[RT/INT]**. Looking at this negatively, we can see that her rating scale is only half-complete **[FN]** and she admits that if she were asked why she assigned certain criteria, she would not know how to answer **[INT/RT]**. This probably seems like heresy to some assessors. However, Rachel can use this coping strategy because she has a good grasp of what Pre-sessional 3 students can achieve at the different levels:

> **[RT]** 'It's like when I put there, 'B/C' and I've sort of partially filled this in I've got an idea, a global idea about where this is and you know I could be minded to fill it in so it showed that, or I could be minded to fill it in so it shows that, and I'm conscious in this of my role as an educator as much as my role as an assessor'

Rachel justifies this holistic approach by perceiving the ACP assessment as equally formative as summative **[INT]**. This explains why she gives an 'A' for Grace's language, the only rater to ascribe an A score. According to Rachel, it is important to boost the performance's strengths to mitigate the weaknesses:

> **[RT]** 'You're partly wanting to offer encouragement and recognition of things attempted rather than things achieved so that mitigates in the direction of not being too harsh'

Rachel is the only rater to make it explicit that Pre-sessional 3 is a transitional course and therefore she appears more lenient to Grace's lack of criticality. She explains this by observing that criticality is something that will be developed more extensively on Pre-sessional 4 **[RT]**: particularly in the summative assessment.

This contrasts with Sarah, who takes the more traditional approach to assessment which clearly separates the teachers' role from the assessor's role **[RT]**. However, they are not diametrically opposed because Sarah also recognised the developmental needs of the students when she stressed how important it was to assess language fairly and to assess whether students ‘were saying what they wanted to say’ as a counterbalance to the generally low criticality scores across the cohort **[INT]**.

I would now like to briefly look at Ruth and Leah’s interview responses to see if any similar patterns emerged from their interviews. Rebecca did not discuss any strategy-use in her interview and this will be explained later.

Like Rachel, Ruth is well-experienced in rating ACPs and observes that, as a result, she can quickly get a sense of what the final score is going to be **[INT]**. This global impression acts as a ‘guide’ for her rating, but she does not fit that global impression to the criteria **[INT]**. She uses the criteria to revise and ‘fine-tune’ the global impression **[INT]**. Furthermore, her overall score is not finalised until all the criteria have been added up **[INT]**.

For Leah, a ‘gut reaction’ to a performance is also a useful strategy **[INT]**. From her IELTS examiner experience, she would often ask herself, ‘Is this student a 6 or a 6.5?’ and she had a ‘feeling for this distinction’ which may override the descriptors on the rating scale **[INT]**. Faced with the vagueness and quantity of the EAP Unit’s descriptors, she does not get ‘bogged down’ in the details, but employs a combination of global impression and reviewing that impression by analysing the finer details **[INT]**. This may explain why she left question marks by so many of the criteria, electing to negotiate these with her co-marker **[ST]**.

In summary, it would be over-simplistic to say that Sarah and Rachel fit Pollitt and Murray’s (1996) distinction of the analytic and synthetic rater. They each have elements of both identities, although Sarah did seem to be more focused on the distinctions between the criteria and Rachel’s elicitation seemed to highlight the global impression. There was also a salient difference in how they approached the rating process, with Sarah more focused on applying the criteria fairly and Rachel more focused on the developmental aspect. There also seemed to be a pattern that several of these experienced raters use which is an interactive process of mentally applying a global, overall score at the beginning of performance and then checking this, using bottom-up processes, against the criteria or against the details of the unfolding performance. In Rachel and Leah’s case this seems to be a direct result of over-detailed descriptors, but for Ruth this is not made explicit: it could be a general strategy she uses with all speaking assessment.

5. Limitations

A salient limitation of the research is that several raters felt that the verbal report did not exactly replicate authentic rating conditions. This is because the procedure gave them an opportunity to spot features of performance they missed the first time, and, being in control of playback, they had time to notice more features. This is mentioned in the interviews of Rachel and Rebecca. Rachel elaborates that her approach would have been the same in both viewings, but her scores might have been different because she might have focused on different aspects of performance in both viewings. When I tried to triangulate some of the decision-making processes which emerged from Rebecca's transcript, Rebecca said that processes such as predicting, comparing what Grace was saying to what she had previously said, or decoding written and aural information simultaneously may have had more to do with the methodology itself, rather than her personal strategy. I therefore kept these comments in mind when I was inferring patterns from my results.

Sarah, Leah and Ruth, however, were positive about their experiences of the verbal report, with Sarah saying that her score would have remained the same in both viewings and Ruth observing that, although the methodology may have had 'a little effect on her rating', she would have made the same judgements. Furthermore, Ruth observed that the verbal report reflected the internal dialogue that normally occurs in her head when she is rating. This mixture of responses emphasises the importance of triangulating verbal report data with other research methods.

The second major limitation is the subjectivity of the coding, which was mentioned in the methodology section. During the data analysis, it was made clear to me how difficult it is to 'train' a classmate to understand my research topic and the rating scale. It was also imperative to have more time not only to co-code a significant amount of the data (Gass and Mackey, 2000), but also discuss similarities and differences between the two code schemes. From this experience, I concluded that this kind of research requires at least two dedicated researchers who fully understand the research context.

6. Conclusion

This study has shown that the raters at the EAP Unit overwhelmingly heeded construct relevant information (Fulcher, 2003), and were extremely adept at thinking in terms of the rating scale. I suggested that this was due to the amount of rater training that they had undergone. In their evaluations the raters also showed a strong orientation towards criticality, which reflects the central position which critical thinking holds in EAP assessment. Although many of the raters were positive about the equal division of genre, criticality and language in the rating scale, and the range of authentic academic skills that are tested in the construct, they also found the rating scale difficult to manage. The large number of criteria, the level of detail and the use of vague qualifiers were common complaints from the raters. The research also found that raters seemed to be using different coping strategies to deal with this problem. These included focusing on the easier-to-score categories first (genre and language), focusing on the criteria that the rater personally thought were the most salient, scoring as many criteria as possible within one sub-category, or relying on a co-rater to explain ambiguous descriptors or negotiate a score.

The verbal report also found variation in how the criteria ‘Use of Sources’, ‘Sources fully and correctly referenced’, ‘Linking the presentation to the conference theme’ and ‘The reporting of other people’s ideas and stance’ were being interpreted. Because of these problems, I suggested an empirical approach to rating scale design to better capture observable behaviours of students at different levels (Luoma, 2004). However, to do this, the EAP Unit would need to consider whether this gain in the validity and manageability of the rating scale is worth the practical costs of time, labour and money to develop (Knoch, 2009). The EAP Unit might conclude that the flaws in the rating scale are offset by the extra pair of eyes and the ability to negotiate the meaning of a descriptor, which a co-marker provides. Several raters also implied that the range of criteria serves an important diagnostic function.

If the EAP Unit were to redesign the rating scale using intuitive methods, or use a combination of methods, which is the preferred approach of Luoma (2004), I would suggest

7 to 10 criteria in an analytic scale. Given the way that many of the criticality judgements on depth of analysis, use of sources and recognising alternative arguments seemed to merge into one in the verbal reports, I would suggest one global criterion for criticality, but I would give it more weight than the other criteria. This would reflect the strong orientation towards criticality that the raters both implicitly showed in their verbal reports and reinforced in their interviews.

This study showed several raters using the strategy of forming a global score early on in the performance, and then checking that score against the criteria, using bottom-up processing. They seemed to be able to do this due to their experience of rating the ACP genre. However, it was not clear whether this was a coping strategy for this cognitively-demanding assessment or whether raters used this strategy to rate all their speaking assessments. It would be interesting to do further verbal reports on these raters in other assessments to see if they used the same combination of global impression and bottom-up revising.

I could not find a clearcut division between analytical raters and synthetic raters (Pollitt and Murray, 1996). Both Sarah and Rachel formed global impressions of performances, but were open to adapting these judgements as the performance unfolded. Nor was there a clearcut difference in terms of philosophy. Rachel explicitly referred to the developmental aspect of the ACP assessment, whereas Sarah seemed more objective and assessment-focused. However, both Rachel and Sarah focused on language as a way of mitigating a low criticality score. The difference was that Rachel did so explicitly in the verbal report, but Sarah only referred to it in the interview.

Because there are no precedents for a verbal report on rating the ACP genre, there are several salient points which emerged from this research which could be useful to other researchers. On a practical level, the video proved to be a successful training method for this type of report and I would recommend it for L1-speaking participants. The code scheme could also be adapted by other researchers, given that there are so few published code schemes for rater strategies in speaking assessment.

I had originally planned to do several verbal reports with each rater, but I found that I could not edit the presentation down to a ‘manageable chunk’ of 8 to 10 minutes without losing vital parts of the content. The criticality, as the raters confirmed, cannot be rated until a rater has seen the whole performance. Consequently, I would conclude that a short interview or paired speaking task of 3 to 5 minutes is more suited to the verbal report methodology. This is a useful finding for any researcher thinking of doing a similar procedure with the ACP genre. I also concluded that a verbal report study needs more than one dedicated researcher to reliably co-code the data. Finally, I demonstrated how vital it is to triangulate verbal report data with another research method, particularly for investigating rater strategy use. When used on its own, I feel the verbal report is an invalid tool for investigating rater strategies, both because it warps the rating process to a certain extent, and because certain participants are more likely to reflect on their decision-making processes than others. However, it can be a good way to notice a few emerging tendencies which can be developed further, or challenged, by an interview with the rater.

7. Bibliography

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

British Association of Lecturers in English for Academic Purposes (BALEAP) (2008) *Competency Framework for Teachers of English for Academic Purposes.* Available at: https://www.baleap.org/wp-content/uploads/2016/04/teap-competency-framework.pdf (Accessed: 9 August 2016).

Brown, H.D. *Language Assessment: Principles and Classroom Practices.* New York: Longman.

Brown, A. (2000) 'An investigation of the rating process in the IELTS oral interview', in Tulloh, R. (ed.) *IELTS Research Reports 3.* Melbourne: IELT Australia, pp.49 – 84.

Brown, A., Iwashita, N. and McNamara, T. (2005) *An Examination of Rater Orientations and Test-taker Performance on English for Academic Purposes Speaking Tasks (TOEFL Monograph Series Number MS29).* Princeton, New Jersey: Educational Testing Services.

Brown, J.D. and Rodgers, T.S. (2002) *Doing Second Language Research.* Oxford: Oxford University Press.

Cassily, C. and Ventola, E. (2002) 'A multi-semiotic genre: the conference slide show', in Ventola, E., Shalom, C. and Thompson, S. (eds.) *The Language of Conferencing.* Frankfurt: Peter Lang, pp.169 - 209.

Chapelle, C.A. and Brindley, G. (2010) 'Assessment', in Schmitt, N. (ed.) *An Introduction to Applied Linguistics.* 2nd edn. Abingdon: Routledge, pp.247 – 268.

Cotton, F. (2010) 'Critical thinking and evaluative language use in academic writing: a comparative cross-cultural study' in Blue. G. (ed.) *Developing Academic Literacy.* New York: Peter Lang. pp.73 – 87.

Cottrell, S. (2011) *Critical Thinking Skills: Developing Effective Analysis and Argument. Basingstoke: Palgrave MacMillan.*

Cumming, A., Kantor, R. and Powers, D.E. (2002) ‘Decision-making while rating ESL/EFL writing tasks: a descriptive framework’, *The Modern Language Journal,* 86(1), pp.67-96.

De Chazal, E. (2013) *English for Academic Purposes.* Oxford: Oxford University Press.

Dörnyei, Z. (2007) *Research Methods in Applied Linguistics.* Oxford: Oxford University Press.

Forey, G. and Feng, D. (2016) ‘Interpersonal meaning and audience engagement in academic presentations: a multimodal discourse analysis perspective’, in Hyland, K. and Shaw, P. (eds.) *The Routledge Handbook of English for Academic Purposes.* Abingdon: Routledge, pp.416 – 431.

Frobert-Adamo, M. (2002) ‘Humour in oral presentations: what’s the joke?’ in Ventola, E., Shalom, C. and Thompson, S. (eds.) *The Language of Conferencing.* Frankfurt: Peter Lang, pp.211 - 225.

Fulcher, G. (1996) ‘Does thick description lead to smart tests? A data-based approach to rating scale construction’, *Language Testing,* 13(2), pp.208 – 238.

Fulcher, G. (2003) *Testing Second Language Speaking.* Harlow: Pearson Education Limited.

Gass, S.M. and Mackey, A. (2000) *Stimulated Recall Methodology in Second Language Research.* Mahwah, New Jersey: L. Erlbaum.

Gebril, A. and Plakans, L. (2014) ‘Assembling validity evidence for assessing writing: rater reactions to integrated tasks’, *Assessing Writing,* 21, pp.56 – 73.

Green, A. (1998) *Studies in Language Testing 5: Verbal Protocol Analysis in Language Testing Research.* Cambridge: Cambridge University Press.

Halliday, M.A.K. (1970) ‘Language Structure and Language Function’, in Lyons, J. (ed.) *New Horizons in Linguistics.* Middlesex: Penguin, pp. 140 – 165.

Halliday, M.A.K. and Hasan, R. (1989) *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective.* Oxford: Oxford University Press.

Hood, S. and Forey, G. (2005) ‘Introducing a conference paper: getting interpersonal with your audience’, *Journal of English for Academic Purposes,* 4, pp.291 – 306.

Hutchby, I. and Wooffitt, R. (2008) *Conversation Analysis.* 2nd Edn. Cambridge: Polity Press.

Hyland, K. (2005) ‘Stance and engagement: a model of interaction in academic discourse’, D*iscourse Studies*, 7(2), pp.173 – 192.

Knoch, U. (2009) ‘Diagnostic assessment of writing: a comparison of two rating scales’, *Language Testing,* 26(2), p.275 - 304.

Lumley, T. and Brown, A. (2005) ‘Research Methods in Language Testing’, in Hinkel, E. (ed.) *Handbook of Research in Second Language Teaching and Learning.* Mahwah, New Jersey: L. Erlbaum, pp.833 – 857.

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

Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands (2016) *Elan* (Version 4.9.4.) [Computer program]. Available at: https://tla.mpi.nl/tools/tla-tools/elan/citing_elan/ (Accessed: 20 August 2016).

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

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

McKay, S.L. (2009) ‘Introspective Techniques’ in Heigham, J. and Croker, R.A. (eds.) *Qualitative Research in Applied Linguistics: a Practical Introduction.* Basingstoke: Palgrave MacMillan, pp.220 – 241.

McNamara, T. (1996) *Measuring Second Language Performance.* Harlow: Addison Wesley Longman Ltd.

Morell, T. (2015) ‘International conference paper presentations: a multimodal analysis to determine effectiveness’, *English for Specific Purposes,* 37, pp.137 – 150.

Nesi, H. and Gardner, S. (2012) *Genre across the Disciplines: Student Writing in Higher Education.* Cambridge Cambridge University Press.

Orr, M. (2002) ‘The FCE speaking test: using rater reports to help interpret test scores’, *System,* 30, pp.143 – 154.

Pollitt, A. and Murray, N.L. (1996) ‘What raters really pay attention to’, in Milanovic, M. and Saville, N. (eds.) *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium.* Cambridge: Cambridge University Press, pp.74 – 91.

Polushkina, T. (2013) *IELTS Speaking Part 2: a Well-known Person Band (6.5).* Available at: https://www.youtube.com/watch?v=_0DYAYAV6Xk (Accessed: 9 August 2016).

Rice, R., Boysen, A. and Stetler, L. (2004) ‘Assessing oral presentations and writing skills’, *Proceedings of Professional Communication Conference, IPCC, 2004*. 147 – 150. September – October. doi: 10.1109/IPCC.2004.1375289.

Richards, J.C. and Schmidt, R. (2010) *Longman Dictionary of Language Teaching and Applied Linguistics*. 4th edn. Harlow: Longman.

Richards, K. (2003) *Qualitative Inquiry in TESOL.* Basingstoke: Palgrave Macmillan.

Shalom, C. (2002) ‘The academic conference: a forum for enacting genre knowledge’, in Ventola, E., Shalom, C. and Thompson, S. (eds.) *The Language of Conferencing.* Frankfurt: Peter Lang, pp.51 – 68.

Upshur, J.A. and Turner, C.E. (1995) ‘Constructing Rating Scales for Second Language Tests’, *ELT Journal,* 49(1), pp.3 – 12.

Ventola, E. (2002) Why and what kind of focus on conference presentations?’, in Ventola, E., Shalom, C. and Thompson, S. (eds.) *The Language of Conferencing.* Frankfurt: Peter Lang, pp.15 – 50.

Weir, C.J. (2005) *Language Testing and Validation: an Evidence-based Approach.* Basingstoke: Palgrave MacMillan.

Yang, W. (2014a) ‘Strategies, interaction and stance in conference language: ESP presentations made by non-native English speakers’, *Taiwan International ESP Journal,* 6(2), pp.26 – 55.

Yang, W. (2014b) ‘Stance and engagement: a corpus-based analysis of academic spoken discourse across science domains’, *Professional Communication Knowledge Management Cognition,* 5(1), pp.62 – 78.

Zappa-Hollman, S. (2007) ‘Academic Presentations across post-secondary contexts: the discourse socialisation of non-native English speakers’, *The Canadian Modern Language Review,* 63(4), pp.455 – 485.

Zareva, A. (2013) ‘Self-mention and the projection of multiple identity roles in TESOL graduate students’ presentations: the influence of the written academic genres’, *English for Specific Purposes,* 32, pp.72 – 83.