Teacher Rating of Class Essays Written by Students of English as a Second Language: A Qualitative Study of Criteria and Process

Manal Saleh Alghannam

A thesis submitted for the degree of Doctor of Philosophy in

Applied Linguistics

Department of Languages and Linguistics

University of Essex

October, 2017

DEDICATION

To my entire universe

To the secure place for my sails

To my hope

To the ones whom I build my life from

To the eyes which I see with

To my pillars of strength

To my parents for all I am

ABSTRACT

This study is concerned with a neglected aspect of the study of L2 English writing: the processes which teachers engage in when rating essays written by their own students for class practice, not exams, with no imposed rating/assessment scheme. It draws on writing assessment process research literature, although, apart from Huot (1993) and Wolfe et al. (1998), most work has been done on scoring writing in exam conditions using a set scoring rubric, where all raters rate the same essays. Eight research questions were answered from data gathered from six teachers, with a wide range of relevant training, but all teaching university pre-sessional or equivalent classes. Instruments used were general interviews, think aloud reports while rating their own students' essays, and follow up immediate retrospective interviews. Extensive qualitative coding was undertaken using NVivo.

It was found that the teachers did not vary much in the core features that they claimed to recognise in general as typical of 'good writing', but varied more in what criteria they highlighted in practice when rating essays, though all used a form of analytic rating. Two thirds of the separate criteria coded were used by all the teachers but there were differences in preference for higher versus lower level criteria. Teachers also differed a great deal in the scales they used to sum up their evaluations, ranging from IELTS scores to just evaluative adjectives, and most claimed to use personal criteria, with concern for the consequential pedagogical value of their rating for the students more than achieving a test-like reliable score.

A wide range of information sources was used to support and justify the rating decisions made, beyond the essay text, including background information about the writer and classmates, and teacher prior instruction. Teacher comments also evidenced concern with issues arguably not central to rating itself but rather exploring implications for the teacher and writer. Similar to Cumming et al. (2002), three broad stages of the rating process were identified: reading and exploiting information such as the writer's name and the task prompt as well as perhaps skimming the text; reading and rereading parts of the essay, associated with interpretation and judgment; achievement of a summary judgment. In detail, however, each teacher had their own individual style of reading and of choice and use of criteria.

ACKNOWLEDGMENTS

بسم الله الرحمن الرحيم

الحمد لله الذي بنعمته تتم الصالحات

I thank Almighty God (*Allah*) for His graces on me and for giving me the strength and perseverance during this journey. I offer sincere thanks to the many people who have supported me throughout these years and who have provided me with all the help and guidance in completing my research.

First I would like to thank Dr. Julian Good, my supervisor, whose support, guidance, and encouragement helped me throughout the dissertation process. His patience and quite character teach me how to relax. I wish also to sincerely thank Professor. Roger Hawkins, the chairman of my board for his role in helping this work comes to fruition and for his dedication to myself. I feel as though I owe Professor. Hawkins pure thanks and deep appreciation not only for guidance and support, but for his patience, forbearance, unflagging support and his ever-open door. I would like also to thank Christina Gkonou, my advisor for her support and deep understanding.

I would like also to acknowledge the following;

- My parents, who raised me with love, educated me well, and have always supported, encouraged and believed in me in all my endeavours and who are always with me, in happiness and sadness.
- Ahmad Alsamanni, my husband, without whom this effort would have been worth nothing. Your love, support and constant patience taught me so much about sacrifice, discipline and compromise, even if there were times when you said “I told you so”.
- Sulaiman, Rafah and Fulwah, my children, who spent many days waiting patiently by the window for me to come home from the university. I am deeply sorry for the time we spent apart.
- Sharifah Alsamanni, my mother-in-law, who was always sympathetic and count the days for my coming back. Thank you for your prayers.
- My dearest brother Abdulaziz Alghannam, who is always with me and always present whenever I need him. Love you my childhood companion.

- My uncles Abdulrahaman & Abdualaziz Alsamanni for their support and guidance.

I would like to thank the following for their support in so many different ways;Lots of thanks to my friend Shamas Alhamed, Dr. Nada Alkhatib, Dr. Wafa Alsafi and a special thank must go to my colleague Dr. Deema Alammaar, for helping me over the final hurdle, who inspired my final efforts despite the enormous work pressures we were facing together.

Finally, my deepest gratitude goes to all my friends at home and here in Colchester. Namely: Dr. Woroud Melhem, Mayson Khoja, Arwa Basabbrain, Samah Felemban, Huda Altaisan, Enas Jambi, Nora Alkhamees, Shrooq Alhutoi, Faten Adas, Heba Haridi, Halah Almasrani and Faye Abu Abat. Thanks to many other family, friends, and colleagues who are too numerous to mention.

LIST OF ABBBREVIATIONS

CEFR	**Common European Framework of Reference for Languages**
EFL	**English as a foreign language**
ESL	**English as a second language**
EELP	**Essex English language program**
CESC	**Colchester English study centre**
L1	**First language**
L2	**Second language**
IELTS	**International English language testing system**
TEFL	**Teaching English as a foreign language**
PGCE	**Postgraduate certification of Education**
IA	**International Academy in Essex University**
CELTA	**Certificate in teaching English to speakers of other languages**
DELTA	**Diploma in teaching English to speakers of other languages**
TESOL	**Teaching English to speakers of other languages**

TABLE OF CONTENTS

DEDICATION i
ABSTRACT ii
ACKNOWLEDGMENTS iv
LIST OF ABBBREVIATIONS vi
TABLE OF CONTENTS vii
LIST OF TABLES xi
LIST OF FIGURES xii

Chapter 1 INTRODUCTION 1
1.1 The importance of evaluation of classroom writing 1
1.2 The need for research into the role of the rater 4
1.3 Aims of the study 9
1.4 Scope and limitations of the study 10
1.5 Significance of the study 12
1.6 Structure of the thesis 14

Chapter 2 LITERATURE REVIEW 15
2.1 Introduction 15
2.2 Teacher cognition, beliefs and practices concerning writing assessment 16
2.2.1 Definition of cognition and its related terms 16
2.2.2 Studies of teacher beliefs about assessment 20
2.2.3 Research on teacher beliefs/practices in relation to feedback on writing 23
2.3 Factors affecting the rating of ESL students' writing 26
2.3.1 Overview of the factors 26
2.3.2 The need to understand the factors affecting the rating process 32
2.3.3 The nature of the rating scale and criteria as a factor in rating 35
2.3.3.1 Definition of a rating scale 35
2.3.3.2 Alternatives to rating scales 36
2.3.3.3 Types of rating scale: overview 37
2.3.3.3.1 Holistic scales 38
2.3.3.3.2 Primary and multiple trait scales 40
2.3.3.3.3 Analytic scales 42
2.3.3.3.4 Holistic vs analytic scoring 43
2.3.4 The nature of the rater as a factor in rating 47
2.3.4.1 Rater professional background 47
2.3.4.2 Rater cultural background 48
2.3.4.3 Rater professional experience of rating 48
2.3.4.4 Rater linguistic background 49
2.3.4.5 Rater reading/rating style 50
2.3.4.6 Rater preferred scale and criteria 51
2.3.4.7 Rater training 52
2.4 The rating process 53
2.4.1 Key empirical studies of the strategies in the process of rating English writing 55
2.4.1.1 Huot (1988) 55

2.4.1.2 Cumming (1990) 58
2.4.1.3 Vaughan (1991) 62
2.4.1.4 Weigle (1994) 67
2.4.1.5 Erdosy (2000) 70
2.4.1.6 Sakyi (2000) 72
2.4.1.7 Cumming et al. (2002)'s taxonomy of the scoring process 73
2.4.2 Comprehensive models of the rating process as a sequence 77
2.4.2.1 Milanovic, Saville and Shuhong's (1996) framework of the scoring process 77
2.4.2.2 Wolfe's (1997, 2006) framework of the scoring process 81
2.4.2.3 Lumley's (2000, 2005) model of the scoring process 85
2.5 Instruments used to find out about the essay rating process in mother tongue and ESL/EFL contexts 90
2.6 Chapter conclusion 91

Chapter 3 STUDY DESIGN AND ANALYSIS 95
3.1 Introduction 95
3.2 Overview of the study 95
3.2.1 Focus of the study 95
3.2.2 Research approach 97
3.2.3 Research questions in relation to instruments and data 100
3.3 Teacher/rater participants 102
3.3.1 Rationale of selection of teachers/raters 102
3.3.2 Characteristics of raters selected 104
3.3.2.1 Educational background and L1 104
3.3.2.2 Teaching experience and training 105
3.3.2.3 Gender 106
3.4 Composition scripts to be rated 106
3.4.1 Rationale of selection of scripts 106
3.4.2 Selection of scripts and their characteristics 107
3.5 The instrumentation 109
3.5.1 Introduction to the instrumentation 109
3.5.2 General interview 109
3.5.3 Think aloud reporting 111
3.5.3.1 Nature of think aloud reporting and its advantages 111
3.5.3.2 Criticisms of think aloud reporting 112
3.5.3.3 Use of think aloud in previous rating studies 114
3.5.4 Immediate retrospective interviews 116
3.5.4.1 Nature of retrospective interviews and their advantages 116
3.5.4.2 Rationale for use of interviews in the study 117
3.6 Data collection materials and procedures 118
3.6.1 Overview of the procedure 118
3.6.2 Research ethics 119
3.6.3 The general interview 119
3.6.4 The TA training / warm-up 120
3.6.5 The TA verbal reporting 121
3.6.5.1 The verbal report instructions 121
3.6.5.2 The rating task 124

3.6.5.3 Presence of the researcher during TA data collection 125
3.6.6 The retrospective interviews 126
3.7 Piloting the study 128
3.7.1 Conduct of the Pilot 128
3.7.2 Benefits of the Pilot 128
3.8 Validity and reliability of the data collection 131
3.8.1 Validity of the data gathering 132
3.8.2 Reliability of the data gathering 133
3.9 Qualitative data transcription and analysis 134
3.9.1 Initially preparing and organizing the data 135
3.9.2 Transcribing the data 135
3.9.3 Approach to segmentation and coding 139
3.9.4 Initial exploration of part of the TA data 140
3.9.5 The preliminary segmentation of all the TA data prior to full coding 142
3.9.6 Transferring the data into qualitative data analysis software 149
3.9.7 Starting point for the full coding scheme 150
3.9.8 The first cycle of full TA and retrospective interview coding 151
3.9.9 Reviewing the list of codes after the first cycle of coding 156
3.9.10 The second full cycle of coding and its review 159
3.9.11 The final list of codes 160
3.10 Quantitative analysis 165
3.11 Validity and reliability of the data analysis 165

Chapter 4 Results and Discussion 167
4.1 Introduction 167
4.2 Teachers' general perception of 'good writing' 168
4.3 Rating scale and general criteria of the teachers, and their sources 175
4.3.1 Scales and criteria in relation to their sources 175
4.3.2 Scales and criteria in relation to the pedagogical function of the assessment 179
4.3.3 Variability within teachers in criteria used and their weighting 183
4.4 Teachers' training in writing assessment and views on such training 187
4.4.1 Relevant training received 188
4.4.2 Attitudes to training 190
4.5 Teachers' actual rating criteria reported used in their assessment of essay qualities 195
4.5.1 Frequencies of mention of criteria 196
4.5.2 Wording of evaluation of essays using the criteria 201
4.5.3 Summarising or combining ratings/scores for separate criteria into an overall score/rating 206
4.6 Non-text-based support for rating or scoring judgements 210
4.6.1 Retrospective considerations 211
4.6.2 Prospective considerations 215
4.6.3 Consideration of rating impact on the writer 217
4.7 Wider aspects talked about 220
4.7.1 Suggested reasons for student performance 221

4.7.2 Implications for teachers and pedagogy 223
4.7.3 Feedback to the writer 226
4.7.4 Personal reaction to ideas or to students 229
4.8 Sequences and stages of rating related activity 231
4.8.1 Stage 1: Reading information for activating the relevant background information 233
4.8.1.1 Identifying task prompt 234
4.8.1.2 Identifying student's name 235
4.8.1.3 Initial impression 239
4.8.2 Stage 2: Reading the essay, engaging in interpretation strategies, while exerting certain judgment strategies (applying criteria) 241
4.8.2.1 Reading behaviour 241
4.8.2.2 Interpretation and judgment 244
4.8.3 Stage 3: Overall impression while summarising and reinterpreting judgments 247
4.9 Individual rater styles 250
4.9.1 Reading for errors then reading for the criteria 251
4.9.2 Reading twice with "three-category" focus 253
4.9.3 Two full reads and focus on quantification of genre related features 254
4.9.4 One detailed read, focus on feedback 256
4.9.5 Much reading and rereading, with varied focus depending upon teacher's expectation and the essay itself 257
4.9.6 One read, with high level criteria focus 258
4.10 Effect of teacher background 260
4.11 Chapter conclusion 262
4.11.1 Bob 262
4.11.2 Gena 264
4.11.3 James 267
4.11.4 John 269
4.11.5 Sam 271
4.11.6 Zain 273
Chapter 5 Conclusion 278
5.1 Introduction 278
5.2 Summary of findings 278
5.3 Significance and implications 282
5.4 Limitations 292
5.5 Future research 295
REFERENCES 298
APPENDICES 315

LIST OF TABLES

Table 2-1'Personal comment' categories (Huot1988; and Pula & Huot 1993) (adapted from Huot 1988; and Table 1. Pula & Huot 1993, p.244) 57
Table 2-2 Judgment behaviours with self-control focus (Cumming 1990, p.37)..... 60
Table 2-3 Categories of comments made by raters in (Vaughan's 1991 study, p. 116)... 65
Table 2-4 Descriptive framework of decision-making behaviors while rating TOEFL writing tasks (Cumming et al. 2002, p.88) 75
Table 2-5 Model of the stages in the rating sequence (Lumley 2002, p. 255) 86
Table 3-1 RQs and the main data used to answer them 101
Table 3-2 Demographic information about participants 105
Table 3-3 Examples of decisions made during the first cycle of full coding..... 152
Table 4-1Teacher ratings of seven aspects of 'good writing' 169
Table 4-2 Supplementary features of 'good writing' mentioned by teachers 170
Table 4-3 Sources of criteria that the teachers employed 176
Table 4-4 Training courses that teachers received 188
Table 4-5 Teacher positive comments on the usefulness of training..... 191
Table 4-6 Frequencies of references to different criteria in the TA..... 197
Table 4-7 Criteria in order of overall frequency of mention 200
Table 4-8 General characteristics that determined teachers' judgments of the quality of the written compositions..... 202
Table 4-9 Summary of teachers' good and bad ratings in some students' compositions 204
Table 4-10 Information sources used by raters to support or justify ratings that they make, and for other purposes 220
Table 4-11 Frequency of reading task prompt..... 235
Table 4-12 Comments about student's level or background at the first stage..... 236
Table 4-13 The effect of prior knowledge of the writer's name and background level in the process of evaluation 238
Table 4-14 Frequency of reading behaviour during Stage 2 reading. 243

LIST OF FIGURES

Figure 2-1 Factors in performance assessment (adapted from McNamara, 1996, p. 9) . 28

Figure 2-2 A model of the decision-making process in holistic composition marking (Milanovic et al., 1996, p.95) 79

Figure 2-3 Model of scorer cognition (Wolfe, 1997, p. 89) 83

Figure 2-4 A detailed model of the rating process (Lumley, 2000, p. 289) 90

Figure 3-1 The room setting during think-aloud and retrospective interview..... 130

Figure 4-1 Sequence of the rating process while evaluating students' compositions. .. 233

CHAPTER 1 INTRODUCTION

1.1 The importance of evaluation of classroom writing

In the context of second language learning, teachers are constantly assessing their students' performance, noting their writing, reading, pronunciation, their use of vocabulary, syntactic rules, and appropriacy of language use (Douglas, 2010). From the teacher saying *very good* to a student speaking in class, to a student gaining an IELTS examination score of 5.5, evaluation in some form is present. Furthermore, these days it is not just the teacher or examiner who evaluates, but also the student's peers in a group work activity, or the software the student uses to learn independently online.

Such evaluation goes under many names, including assessment, grading, scoring, marking, rating etc. Furthermore, it can have any of a variety of purposes, not just to provide overt recognition of a proficiency level attained, but oftentimes to reward, warn, motivate, inform and so forth.

Within this broad spectrum of evaluative activity, some areas have been studied in greater depth than others. In particular, the current work draws on our observation from the literature (see further Chapter 2), that while formal assessment by examiners or testers of work done by learners under examination conditions has received an enormous amount of attention, both by researchers and examining boards themselves, less attention has been paid to the less formal types of evaluation which occur all the time in teaching contexts as part of the day to day teaching and learning of English (Cumming et al., 2001, 2002)

In particular, the present research focuses on writing, as a key language skill where this inconsistency of research attention may be seen. As our literature review in chapter 2 will show, much research attention has been lavished on all kinds of issues concerning the scoring or rating of writing in test conditions. The kinds of scoring systems and criteria that can or should be used, how consistently examiners apply them in the process they follow when scoring scripts, how to ensure the reliability and validity of assessment of writing where extreme accuracy is required, due to the high stakes involved and the consequences for test takers if mistakes occur, all these have been widely studied.

By contrast, what of the vast majority of evaluation of writing performance which occurs on a daily basis in classrooms around the world, simply as part of language practice activities, written communicative class tasks, and so forth? As our review in chapter 2 will again show, it is quite hard to find much attention paid to this. Yet it surely differs in many respects from assessment of writing tests and exams. The teachers doing the assessment may well not be highly trained in any particular system for rating writing. The purpose is not usually to obtain a highly accurate measure of proficiency, but rather to form part of the wider feedback being given to students, to help them improve their writing.

Furthermore, evaluation of this sort, as a key feature of the teaching and learning process which ultimately prepares many students for examinations, is not to be seen as of little importance. Although performance-based assessment has become the method of choice in judging writing ability in English as a Second Language (Kroll, 1998, p. 221), it is still not an easy task for the teacher. It can be seen as a time-consuming and complex activity. As a teacher, I believe that teachers dedicate a large part of their time

to marking in their daily teaching. Hence matters such as what criteria teachers use, what scale of scores, grades or evaluative words they employ, and what process they go through to arrive at a rating have just as much importance to be understood.

It must be noted of course that a great deal of attention has been paid to classroom *feedback* on student writing, whether from teachers, peers or writers themselves working with a checklist (e.g. Cohen and Cavalcanti, 1990; Fathman and Whalley, 1990; Ferris, 1995; Hyland, 2003; Yang et al., 2006; Lee, 2009). Part of such feedback, especially where it comes from a teacher, can indeed be the kind of evaluative rating which we are talking about. Yet once again the focus of attention in the literature has not been on this. The feedback literature rather focuses on types of feedback other than an evaluative rating, such as error correction, communicative comments and so forth, their types, the means by which feedback of these sorts may be given, their effectiveness, student preferences for different feedback types and so forth. The type of evaluative rating which we are concerned with once again often seems to fall just outside the purview of these studies (Leki, 1991).

Yet once again this does not mean that such evaluation is unimportant. The feedback literature indeed often laments, in fact, that the only part of teacher feedback that some students pay any attention to is the evaluative rating or score provided, and not all the informative feedback which the teacher has taken time and trouble to provide on errors and so forth. Despite this recognition of the importance of teacher rating, however, paradoxically this has not led to a great deal of attention being paid in the feedback literature to what criteria teachers use or how precisely they arrive at such a rating.

Hence, the present study aims to contribute to filling this gap in our understanding, with respect to EFL writing. We adopt the term *rating* for the kind of evaluation we are

talking about, since we do not wish to suggest that our study is limited to how teachers award numerical marks, for example, as the word *scoring* would suggest.

1.2 The need for research into the role of the rater

Addressing the topic of writing performance assessment directs our attention to the most significant factor in this process, which is the rater (Lumley, 2002). Raters are at the heart of the rating process, especially of the type we are concerned with, because they can make decisions about what criteria they will focus on, how to adapt a scale, if any is used, how to change scale wording to suit their situations, and how they evaluate the written texts to suit their educational contexts and requirements. Consequently, raters need to keep many things in mind while they assess in order to maintain appropriateness of the rating.

When assessing writing tasks, raters' prejudice may play an important role. Raters' biases towards student performances, their different perceptions of good writing and their culture, mother tongue, academic background and professional experience are all factors which can influence the rating process (Kobayashi, 1992; Wood, 1993; Shi, 2001; Cumming et al., 2002; Erdosy, 2002, 2005). However, such influence arises not only from raters' cultural or disciplinary backgrounds but also varies "according to the types of written genres assessed", as Cumming et al. (2002, p. 68) added. Even time of marking in the day may play a part.

Wood (1993) and McNamara (1996) further postulated that there are many variables involved in rating writing arising from the specific writing being assessed and so the writer. There may be evidence of 'between-writer' variation, or 'within-writer'

variation and even involvement of superficial features of written products such as neatness of handwriting. Recently, writers' nationality has been in focus, as having an effect on writing assessment (Lindsay & Crusan, 2011).

Lumley (2002) further highlights a specific area of concern in the rating process: the superficiality of rating scales of the type commonly used as compared to the complexities that operate in written texts and the subjectivity involved in the interpretation of these scales. Such a mismatch needs to be further examined in a detailed manner so that rating instruments and processes can be fine tuned to minimize unfair practices in the assessment of writing competencies. This again suggests that, where a scale is not even provided, as in our case, a broad range of criteria and interpretations by raters may be revealed.

All the above may apply in our informal assessment context as much as in exam/test related writing evaluation. However, our attention is on factors associated with the rater rather than the students whose writing is rated or their written products, or indeed the impact of an imposed rating scale and criteria.

Considerations such as those just mentioned lead researchers to question the reliability and validity achievable in assessment based on rating, which is necessarily subjective. In formal exam related writing assessment, the literature has shown improvement in the reliability of writing performance assessment over the years through the use of rating scales combined with training (Lumley, 2005). In the last decades, the reliability of formal writing performance assessment has also been strengthened through a variety of approaches such as using a battery of tests (Milanovic et al., 1996, p.92), and better specification of scoring criteria and tasks (Lumley, 2002). However, even in this context of controlled exam related assessment of writing, Milanovic et al. (1996)

highlight the need for more attention to be given to rater behaviour in the marking process as the rater was recognized as "one of the main sources of measuring error in assessing a candidate's performance" in the context of research at Cambridge. They call for a "better understanding of the value, decision-making behaviour and even the idiosyncratic nature of the judgements markers make" (Milanovic et al., 1996, p.92). In informal writing assessment such as we focus on, we may expect to find this even more prevalent.

A number of research studies have contributed to a growing body of literature on the rater in rating process, usually in the context of exam type rating. Cohen (1994) reviewed several studies of both first and second language writing (e.g. Rafoth & Rubin 1984; Hout 1990; McNamara, 1990; Connor & Carrell 1993) and found that raters were likely to focus on mechanics and grammar more than they realize. Some researchers found that raters tend to employ criteria different from those in the guidelines they receive (Cohen 1994; McNamara 1996). In second language assessment contexts, the most important studies of these issues include Cumming (1990), Cumming et al. (2001), Vaughan (1991), Weigle (1994a, 1994b), Zhang (1998): see further chapter 2.

Several studies attest that teachers with different backgrounds will have different perceptions of good writing and thus tend to focus more on some specific features. Shi (2001) investigated how different native speaker English teachers and nonnative speaker English teachers rated their Chinese university students' writing. The teachers were asked to rate writing samples holistically using their own criteria and provide three reasons based on rank of importance to support their judgments. The result showed that though both groups of raters gave similar scores to the writing, they weighted writing features differently in their rating. The native English teachers focused more positively

on content and language while Chinese teachers stressed more negatively the organization and length of the writing.

In Cumming et al.'s (2002) project designed to investigate features of TOEFL essays, rated by experienced raters, they found that the qualities perceived as good writing in the examination context varied among raters. Ten raters were asked to identify the three most significant features of writing from their point of view. It was discovered that the most frequently mentioned features were rhetorical organization (nine raters), expression of ideas (nine raters), accuracy and fluency in English grammar (seven raters) and vocabulary and length of writing (two raters). However, in this study, Cumming stresses the need for further research in order to illuminate better some of the ambiguity associated with rating process.

Chinda's (2009) study, exploring Thai teachers' perspectives on writing assessment practices, discovered that teachers had different attitudes towards criteria and employed them differently. Even though they had central criteria to follow, they applied them in individual ways. However, some teachers tried to follow the criteria, even though they did not agree with them. Some added their own criteria when marking students' work.

Besides the above evidence of rater variability, even when following an imposed scoring system, and calls for more studies of the role of the rater, further studies have repeatedly called for further investigation of the rating process itself (Huot 1993, Paula & Huot 1993). Vaughan (1991), Hamp-Lyons (1991), Wiegle 1994a, b, 1998), Conner-Linton (1995) and Kroll (1998) have all made cases for further exploration in the area of rating process in the ESL context. Despite this body of work, the investigation of the rating process and the basis of raters' decisions is however still regarded as at a preliminary stage (Lumley 2002). This concern echoes Huot's (1990) comments that

> "[. . .] little is known about the way raters arrive at these decisions . . . we have little or no information on what role scoring procedures play in the reading and rating process." (p. 258)

In response to this and other calls for investigating the rater's rating process, in order to understand writing assessment practices more, we deem it to be worth not only exploring raters' perceptions concerning good writing and writing assessment but also investigating how they actually mark their students' writing. As Connor-Linton (1995, p.763) remarks:

> "If we do not know what raters are doing and why they are doing it, then we do not know what their rating means".

Overall, we may conclude that although clearly writer factors, and the nature of the scoring system (where one is imposed), play a role in writing assessment, the need for research into the impact of rater variables and the decision making process of raters themselves is particularly evident. This is true in numerous publications concerning scoring written compositions in the contexts of both English mother tongue (EMT e.g., Charney 1984; Huot 1990; Purves 1992) and English as a second or foreign language (ESL/EFL; e.g., Brindly 1998c; Connor-Linton 1995a; Cumming 1997; Kroll 1998; Raimes 1990). Although the reliability of writing performance assessment in test/exam contexts has been improved over the years, to a substantial extent through the use of better articulated and defined rating scales combined with training, even their researchers have pointed out that the validity of rating has been insufficiently addressed. To make it clear, Cumming *et* al. (2002) say: "a principle criticism has been that the exact nature of the construct they assess remains uncertain" (p.68). This research will

try to fill the gap referred to by such calls, focusing on the situation where in fact no rating/scoring system is imposed.

1.3 Aims of the study

Following from the above, the aim of this study is to continue in the line of investigating the crucial role of the rater and his/her rating process, focussing on an ESL classroom writing context. This study will explore the assessment practices of writing teachers in the UK teaching writing courses to prepare students who are about to study at university through the medium of English. Moreover, the aim of my study is to provide a detailed description of the way six raters rated ESL compositions in the absence of any guidelines (criteria or rating scale), a condition which is not reflective of standard writing test assessment practice, but quite common in informal classroom assessment of practise writing assignment, which, as we stated earlier, is a relatively neglected research area. The approach is inspired by Cumming et al. (2002), which we felt to reveal something of the complexity of the process.

Specifically, this study will address the following questions, which will be more fully justified and explained by Chapter 2:

1. What do ESL writing teachers perceive as good writing?

2. What rating scale and criteria do the teachers claim to typically use when rating student writing, and what are their sources?

3. What training have the teachers received that is relevant to assessing writing and what are their views on such training?

4. What are the most important qualities that the teachers look at in practice when they are rating their students' samples? (criteria used and their weighting).

5. How do the teachers support explain or justify their rating/scoring?

6. What other kinds of comments do the teachers make, beyond those directly related to achieving the rating itself?

7. What common sequence of activity can we find in the rating behavior of the teachers?

8. What distinct rating styles can we detect being used by the teachers?

9. Is there any difference in any of the above between individual teachers, according to their general training, experience of rating writing, prior knowledge of a specific rating system, etc.?

1.4 Scope and limitations of the study

At this point it is suitable to make clear what precisely this study is *not* concerned with. It needs to be distinguished on the one hand from the common kind of study undertaken of the testing of writing and on the other from the common kind of study of teacher classroom feedback on writing.

With respect to the former it differs in that the essays being rated are not written under exam conditions nor for the purpose of testing students' writing ability (achievement or proficiency) but as everyday class practice which forms part of the teaching/learning process. Hence, the purpose of rating is different, and the stakes lower, and, unlike in most tests/exams, the teacher/rater personally chose and set the writing task, and when rating knows who the writer is and has background knowledge of that person.

Furthermore, the teacher/rater is not following any present system of scoring or grading writing imposed by a particular exam or institution.

The above has a number of consequences. For instance, the literature on scoring writing in tests and exams is very concerned with rater reliability and often gets multiple raters to rate the same scripts so as to investigate various factors which might lead to inconsistencies in the scores awarded (Lumley, 2005). This however cannot be a part of our study since we are concerned with the real life classroom writing situation where each teacher sets different assignments to their class, uses their own rating criteria, and it is most unusual for a teacher to rate essays other than those of their own students.

Another preoccupation of the writing testing literature is with how different raters understand and apply the scoring rubric, criteria, scale and so forth which are provided by the exam/test. Clearly once again this is irrelevant to us since the teachers are not all using the same rating system, and may not be using any clearly defined system at all.

A third consequence of the above is that we are not concerned with the actual scores, or ratings in other forms, that individual students actually obtain. Whether this or that student does well, or indeed whether the teacher tends to view the students generally as poor on grammar but good on vocab, is not our concern. Our interest is rather in what criteria they use, how they arrive at an overall rating (if indeed they do), what process they follow, and so forth.

With respect to studies of feedback, the difference of our study is perhaps more subtle. Feedback is indeed associated with the sort of circumstances which are the focus of our attention, since usually in tests and exams scripts are not returned to students and there is no agenda to provide feedback other than the official score. Rich feedback is however

more associated with writing done as part of classroom practice. There is however a distinction between the rating which might be part of that feedback and the rating which we are concerned with. Our study is not primarily concerned with what feedback the teacher gives to the writer, including any score or rating that might be part of that, but rather with the rating that the teacher arrives at for their own purposes, e.g. to help understand how the student is progressing, perhaps to note down for that purpose in their own records, perhaps to guide future teaching and so forth. That is the rating that the current study aims to elicit information about from teachers, and so to illuminate. Importantly, there could be a difference between the teacher's own rating, of this sort, of a student composition, and any rating which they communicate to the student as part of feedback.

There is of course likely to be a connection between the feedback a teacher gives and their private rating of a composition. In particular, it may well be that if organisation is uppermost in their mind as a criterion for arriving at their personal rating of an essay, then their feedback will contain comments written on the script about organisation. This is however also in principle outside our scope, although the data we obtained was such that in the end we found it made no sense to try to exclude this totally. The evidence for the teacher's rating criteria that we will be most concerned with, however, comes from what they tell the researcher in think aloud and interview responses about their thought processes rather than any feedback that they may write on the script.

1.5 Significance of the study

Connor-Linton (1995; p.764) sheds light on the value of research on the rating process, including the following points which are relevant to us:

- If raters are not responding exclusively to the level descriptors of some established scale in making their rating judgments, we need to know what other factors are involved in their rating process.
- It is a prerequisite for principled improvement of rater training.
- It can help to address a number of fairness issues.
- Knowing more about how raters are responding can help in understanding the backwash effects of different rating procedures.
- It can guide us to forms of evaluation that are consistent with teachers' instructional goals and beliefs about the development of writing skills.

In the context of the present study, a better understanding of the strategies used by the teacher raters will provide insights into how similar or different teachers really are from each other, when not rating essays within the strict framework of an imposed assessment system. Additionally, this study will provide useful information for the training of teachers as raters, since we expect strategies to be identified, described, and critiqued in a more concrete manner. Common patterns that may facilitate consistency can be encouraged while the rationale for idiosyncratic behaviour can be closely investigated and appraised for its impact. This will help raters to have a firmer grasp of what should be done especially at problematic stages of the rating process. Ultimately, training in such steps can only lead to a higher level of self awareness, accuracy, and professionalism in the rating process, a central aspect of language teaching.

It is thus expected that this project will provide better insights on informal writing assessment by teachers, which will inform efforts to create a context in which teachers know how to evaluate good and bad writing in a more professional way. The findings of this study, I believe, may be useful for training which could help to raise teachers'

awareness of what is involved in non-test/exam related writing assessment for those who do not understand how to assess appropriately. More widely, focusing as it does on a common but relatively neglected kind of writing assessment, this study also is designed to be of interest to a wide range of people involved in assessing writing including not only teachers, but also curriculum designers and educational policy makers as well as language test developers, TESOL lecturers, and even learners themselves, as well as researchers.

1.6 Structure of the thesis

Chapter 1 has introduced and delimited the topic of this study, suggested what gap it fills and its value, and presented the research questions.

Chapter 2 reviews the relevant literature and provides evidence for the relevance of the research questions which are posed.

Chapter 3 describes the participants in the study, the instruments used, and the detailed steps in the process of qualitative data analysis.

Chapter 4 presents the findings of the study, answering each research question in turn.

Chapter 5 summarises the key findings, and suggests what implications they have. It also reviews the limitations of the study, and suggests useful lines of further research.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

As we described in chapter 1, the focus of our study is on illuminating the rating of writing performed by ordinary teachers, rather than by trained examiners, and performed on class tasks done by their own students for practice, rather than on tasks performed under strict test conditions for assessment. Furthermore, our study concerns rating where the teachers are not supplied with a uniform marking rubric, score scale and criteria to be applied, but left to use whatever procedure, scale and criteria they individually choose.

Our study is concerned both with the kind of writing assessment which teachers actually perform and with what they in general think about the nature of writing assessment and how it should be done. The latter is clearly in the mind of the teacher, and we will of necessity access how the teachers actually perform their writing assessments also via the teachers' minds, through think aloud reporting. Hence below we will first review relevant aspects of the field of teacher cognition (2.2).

Following that, since there has been relatively little investigation of the sort of rating which we are focusing on (whether performed for writing tasks or any other task), our account will heavily draw on, and apply, theory and findings from research on the rating of writing in the context of testing and examining (performance assessment), so as to see what can be learned from it for our study.

Central to this endeavour is an understanding of the complexity of the rating process. For ease of exposition, we break that part of the chapter down into three main areas: the many factors potentially affecting the rating process and the ratings which are awarded to scripts, particularly the scales and criteria used and nature of the rater (2.3); the nature of the rating process itself as it is performed when rating compositions-rater strategies/ practices and models of those (2.4); the research methods that are appropriate for studying these areas (2.5).

Overall then, the aim of this chapter is to review literature relevant to understanding the reasons for our research questions, and providing ideas as to what we might find and what instruments would be suitable to use to answer them.

2.2 Teacher cognition, beliefs and practices concerning writing assessment

2.2.1 Definition of cognition and its related terms

This study is concerned both with the kind of assessment of classroom writing that teachers spontaneously actually engage in, often these days termed part of their 'practices', and with their candid thoughts about such assessment and how it should be conducted, often these days labelled 'beliefs' (Borg, 2003). Hence it may be seen as at least in part falling into the area of research these days often referred to as teacher cognition (Borg, 2006). In an influential article, teacher cognition has been defined "pre- or inservice teachers' self-reflections; beliefs and knowledge about teaching, students, and content; and awareness of problem-solving strategies endemic to classroom teaching." (Kagan, 1990). In our study the teachers will be inservice, and their reflections, beliefs and knowledge will be about assessing writing. Although

definitions such as this typically do not explicitly mention assessment, we can assume that the reference to teaching, students and content was intended to include it.

Some experts have raised the issue whether there is a distinction between belief and knowledge, which are often both referred to under the heading of cognition, as in the definition above. While some, following a constructivist philosophy, would claim that in the end all knowledge is simply the belief of one or more people, others of a more positivist frame of mind would argue that there exists some objective truth in the world, which constitutes knowledge independent of people's beliefs. Those who follow the latter view would then argue that the difference between belief and knowledge is that knowledge is true while a belief is only thought to be true and may or may not be 'actually' true (M. Borg, 2001). In our study, we will however very much follow the constructivist or interpretivist view that, in the area which we are concerned with, i.e. unregulated teacher assessment of student compositions done for practice rather than exam purposes, it is not appropriate to regard there as existing some 'correct' or true way of proceeding which one may then set out to ascertain teachers' knowledge of. Rather it is an activity where palpably different ways of proceeding may be equally valid for different teachers with different essays and classes and current teaching purposes, even teaching within the same context (which in our case will be pre-sessional English courses in Colchester). Hence, we will not distinguish between belief and knowledge.

In fact, it might be argued that it is instead in the area of the testing of writing for exam purposes, which is not our main focus, that there exists some true and agreed 'knowledge' of how assessment should be done, which might or might not accord with teacher beliefs about that sort of assessment. Certainly, as we shall see in later sections below, there have been many claims made about what is the correct way to do it, e.g.

in the debate between proponents of analytic and holistic scoring schemes. However, even in the assessment of writing of that sort it seems premature to claim that there really is just one correct way of doing it which constitutes universally agreed knowledge that raters should possess.

Beliefs specifically have been defined in general as "psychologically held understandings, premises, or propositions about the world that are felt to be true" (Richardson, 1996, p.103), a definition broad enough that what are often called attitudes also fall within the scope of beliefs. In the world of teaching M. Borg (2001, p.186) defined a belief as: "a proposition which may be consciously or unconsciously held, is evaluative in that it is accepted as true by the individual, and is therefore imbued with emotive commitment; further, it serves as a guide to thought and behaviour". This highlights that teachers' beliefs therefore constitute teacher's "mental lives", or what Freeman (2002) describes as the "hidden side" of teaching. If they are unconsciously held then they are even hidden from the teacher herself. Furthermore, it importantly adds the idea that beliefs may affect practices, an idea that has been much explored in research in this field because, if true, it affords research on teacher beliefs a great deal of importance. As Richardson (1996, p.29) puts it "what teachers do is a reflection of what they know and believe." In other words, teachers' knowledge and beliefs provide the schema or underlying framework which guides most of their classroom actions.

For the purposes of my study I need to make a distinction between teacher beliefs about teaching and teacher beliefs about assessment. The former has been a considerable subject of discussion and research, but the latter, my concern, rather less so, and as we saw above were not even explicitly included in Kagan's definition. However, as stressed by Al-Lamki (2009, p.57): "Teachers may hold personal attitudes and beliefs about all

possible aspects of their professional practices, ...” That therefore includes beliefs held about assessment as well as about teaching.

Furthermore, both effects work within a contextual framework of constraints. Just as a teacher's classroom actions may be affected not only by how they believe they should teach, but also by what the prescribed textbook, syllabus or exams require them to do, so also the assessment they engage in may be influenced not only by their beliefs about what type is most suitable, but also what the teaching institution or Ministry imposes on them in terms of required scoring schemes, frequency and topics for assessment, and so forth. Borg (2006, p.275) has emphasized the importance of the teaching setting on both teacher beliefs and their practices related to all aspects of their professional development: “The social, institutional, instructional and physical settings in which teachers’ work have a major impact on their cognitions and practices.”

Contextual factors, as indicated by Kagan (1992), may interact with teachers’ cognitions in two ways: they may either lead to changes in their cognitions and attitudes, which then affect their practices, or they may lead them to change their practices directly without changing the cognitions underlying them. We may expect to find such effects in our study.

Interestingly, in the realm of research on teacher beliefs about teaching, researchers have identified five main characteristics as the ones that have an effect on teachers’ beliefs and attitudes (Calderhead, 1996; Fang, 1996; McGillicuddy-De Lisi & De Lisi, 2001). These characteristics are years of teaching experience, age, grade level, gender, and educational qualification, and we might expect them to affect beliefs about assessment as well. This will be pursued further in relation to assessment in 2.3.

2.2.2 Studies of teacher beliefs about assessment

One has to look quite broadly to find studies of experienced teacher beliefs about assessment, even in general, let alone specifically related to L2 writing.

Typical studies include that by Chen and Bonner (2017) which surveyed preservice and inservice novice teachers' beliefs about grading and about constructivist teaching approaches (US teachers, across a range of subjects). Teachers who endorsed beliefs about grading which the researchers termed "academic enabling" tended also to endorse constructivist approaches to teaching. More relevant to us, qualitative interviews showed that teachers' reasons for their grading-related judgments were not haphazard, but thoughtful, strategic, and analytical. This encourages us that the present study will yield coherent and interesting data from teachers asked to talk about their assessment of writing. Chen and Bonner however did not examine teacher practices when themselves performing assessment. Indeed, it is left vague whether the grading referred to was that done by the teachers in class or more formal testing.

Bonner, Rivera and Chen (2018) went on to survey how far inservice teacher beliefs about assessment, with reference to standards used, matched classroom assessment practices and external tests. This study took place again, however, in an L1/ESL setting in the USA, in secondary schools. The results nevertheless suggested that the teachers' use of standards in teaching, their classroom assessment preferences, and their beliefs about the test-driven system were moderately correlated. This suggests that we may expect to find some consistency between assessment beliefs and practices among experienced teachers in our context.

By contrast, Kinay has done a number of studies involving teacher beliefs about

assessment in an EFL context (Turkey). They have however again been concerned with teachers of all subjects, not just English, let alone specifically EFL writing. Some concern standardized tests (Kinay and Ardiç, 2017). One however (Kinay, 2018) concerns prospective teachers' beliefs about 'authentic' assessment, which would presumably include composition writing as a form of assessment of writing ability. The study reported general approval of this approach to assessment, but did not, as we aim to, get into teacher detailed beliefs about how exactly such assessment should be or actually is performed, whether by the teachers themselves or professional testers.

Moving now to beliefs related to writing, we often find these conducted with only oblique references to assessment. Henderson et al. (2018) for instance, in a study of L1 secondary school teachers, point to the value of the study of beliefs and practices for informing teacher professional development, since in their context they regarded current professional development not to have successfully resulted in student improvement on large-scale writing assessments. However, they were not concerned with teacher beliefs about such writing assessment, only in how their beliefs and practices with respect to teaching writing affected student results when formally assessed. They did, however, operate with a broad enough definition of writing instruction to be able to reveal some differences between teacher beliefs and practices on matters such as formative feedback, which we would regard as close to the sort of assessment that we are concerned with (see 2.2.3).

Other studies have targeted teacher beliefs about writing assessment more directly, but typically the focus has been on the standardised exam-type assessment existing in the context, over which the teacher has no control, not on the kind of informal classroom

assessment that teacher typically operates with. For instance, Troia and Graham (2016) conducted a survey of a random sample of 482 US teachers (grades 3 - 8) concerning, amongst other things, their perception of the Common Core writing and language standards and assessment system adopted by their state. A majority believed that current state writing tests, although more rigorous than previous tests, did not effectively cover important aspects of writing that needed to be assessed, nor accommodate the needs of students with diverse abilities, and required more time than was available to teachers to prepare students properly. Additionally, many teachers believed the professional development they had received was not enough to help them understand the measurement properties of the assessment or how to use test data to identify students' writing needs. While not without interest such studies do not, however, throw direct light on teachers' own writing assessment beliefs and practices when left to themselves.

Finally, Lee (2011) is much closer to my area of interest in that her study concerns an EFL teacher attempting to put her belief into practice concerning the type of classroom writing assessment that is appropriate. She wished to implement formative assessment, designed to assist future learning, rather than summative assessment, designed to quantify past learning. This was a very interesting project, though it differs of course from the current one where it was for practical reasons impossible for the researcher to implement any intervention to change teachers' mode of classroom writing assessment, so as to study the effects of such a change.

In any case, we feel it is in some sense premature to intervene before the facts of a situation are known. We did not feel able to simply assume, without first studying them, that the teachers in the context of our study would inevitably need to change their ways because they would in their day to day assessment of practice compositions be

following the pattern negatively characterised by Lee as: "Teachers dominate the assessment process as testers, while students remain passive testees. Assessment is something teachers "do to" rather than "with" students, mainly for administrative and reporting purposes (i.e. summative)." (p. 99). As in good Action Research (McNiff, 1993), an intervention/action should follow work done first to illuminate the problem which a later action is meant to solve. Our study could be seen as aiming to perform just this illumination.

In my view, that statement by Lee furthermore does not describe the behaviour of writing teachers in all the teaching situations I have been in, and possibly represents a confusion of two types of assessment which this thesis attempts to distinguish. The summative assessment of writing referred to goes on indeed for administrative and other external purposes in most teaching situations but is not by any means always in the teacher's hands, being often professionally or externally set and graded by ministries, English departments or examining boards. The teacher's own assessment applied day to day to writing which students do for practice is not inevitably just a transfer of such summative methods into a situation where a different approach for formative use might be appropriate. My study will show whether my experiential hunch, or Lee's assumption, is the case in our context of study.

2.2.3 Research on teacher beliefs/practices in relation to feedback on writing

Somewhat different from the above is research on teacher feedback on writing (EFL and ESL). There has been a great deal of work done in this area, by notable figures such as Hyland (e.g. Hyland and Hyland, 2006) and Ferris (e.g. Ferris and Hedgcock, 2005). Some of it is in fact concerned with teacher beliefs and/or practices when giving

feedback, although not typically explicitly worded in terms of cognition, beliefs etc. Much of that concerns error correction, which is often a major part of feedback on writing. This part of feedback research is relevant to us since it typically concerns what teachers believe/do in relation to writing done for practice rather than in formal exams, and assessment/evaluation is clearly involved.

However, as indicated in chapter 1, the present study is not exactly a study of feedback on writing. We are studying how teachers assess classroom practice compositions, which may of course be reflected in the feedback they give to the student but is unlikely to be identical. For instance, teachers may arrive at a grade for a composition which they record for their own benefit but do not communicate to the student as part of the feedback. Furthermore, we should note that research on feedback on writing covers some areas that we will omit entirely as too distant from our interest, such as feedback delivered by peers rather than teachers, the training of teachers to give feedback, and the effects, or lack of them, of feedback on student learning (Hyland and Hyland, 2006).

Ferris in a number of works has promoted a particular three step method of giving feedback on writing which she calls 'approach/response/follow-up' (e.g. Ferris and Hedgcock, 2005; Ferris, 2006). This is a prescriptive guide for teachers which is claimed to be research based, though its precise sources in studies of what teachers actually believe about feedback or do when delivering it are unclear. What is notable about it is that the first stage (approach) exhorts the teacher to call to mind their feedback 'philosophy' as Ferris terms it, which is perhaps to be seen as a mixture of their beliefs and their planned practices. It to be done by the teacher answering a set of five questions: why (purpose of the feedback), when (on draft or final version etc.), what (focus on content, organisation, language etc.), who (the feedback giver, teacher or peer), how (structure, amount, mode). Then at the second stage teachers are required

to follow a certain sequence of activities, including: explain their philosophy to the students, use a scoring rubric or checklist, read through the entire paper before making any comments, identify the most important issues, etc.

While we will be ready to find some elements/categories of Ferris' scheme reflected in our data, one big difference is that we are not presenting the task to our participants as a feedback activity but primarily an assessment/rating activity, though they will be free to voice or write feedback to the students if they wish. Secondly, we are suggesting the very least possible to the teachers about what to do, because we want to know what they spontaneously do. Nevertheless, we may expect to hear them mention their 'philosophy', purposes for the assessment, perhaps a rating rubric or at least what they will focus on and why, etc., and they may or may not read the whole script before starting assessing it.

If we turn to accounts of actual research on feedback, it is particularly noticeable that most accounts fail to make much connection between research on feedback and research on rating/assessing student writing. In a comprehensive recent review (Hyland and Hyland, 2006) interestingly it is only in accounts of automated feedback on computer that the latter is mentioned, because some software in fact combines a computer generated rating of essay quality with feedback on the language etc. We however, from experience, argue that, in reality, teachers often combine the two. That is to say that they typically give feedback both as a grade/rating and comments on language, content etc. together.

A related neglected issue is the recognition that teachers operate with at least two perceptions of the text that they are dealing with. They have their personal vision of the qualities, errors etc. of the composition, and the version of that which they deem

appropriate to communicate in the form of feedback comment, error underlining, etc. to the student. That fact has been recognised since as far back as Zamel (1985) but is not routinely built into feedback studies as a dimension worth investigating. There is an aspect of teacher cognition, then, that has been neglected here. This is even more true of a mark/grade or rating. Here again there may be two versions: the one the teacher regards as true, and the one written as part of the feedback on the practice composition, differing perhaps by being higher so as to encourage the weaker student, or lower so as to discourage slacking off by a good student.

Recognition of these neglected areas of teacher cognition has therefore provided the impetus for my study. Having failed to find a great deal of work relevant to our teacher assessment of writing topic in the cognition and feedback literature, we next move to review sources which do focus centrally on assessment/rating issues. These however again prove to be a little off our topic in that they are usually more concerned with assessment in the context of exams and formal tests of writing, rather than classroom writing practice.

2.3 Factors affecting the rating of ESL students' writing

2.3.1 Overview of the factors

This section will outline some studies which have addressed the influences or factors which affect the rating process. Though they have been mainly researched in the context of second language essay exam assessment, they apply equally to rating of any sort, not necessarily for assessment but as part of formative feedback to students. Furthermore, although more often researched for their effect on the product of rating (scores assigned

to essays), they also potentially impact on the process by which the rater arrives at those scores, which is particularly of interest in our study.

First of all, based on reviewing studies in this area, it is useful to identify three broad kinds of factor which may influence the rater's process of rating, and the rating or score that he/she ultimately assigns to a composition. These are for example reflected in a much-quoted figure from McNamara (1996) which schematically represents different factors which affect the final score given to a test-taker in a typical performance assessment context (Figure 2-1). Scoring in a customary fixed-response assessment (e.g. a multiplechoice task) only includes an interaction between candidate and the test instrument, since rating or scoring based on that is automatic. In open response performance assessment, such as composition rating, on the other hand, there are some additional elements, which involve a rater or judge to assess a sample of performance (e.g. an essay) through a scale or other kind of scoring schedule provided (Weigle 2002).

The additional interactive components of rater and rating scale, which mediate the scoring of writing (and speaking) performance, have opened new areas of investigation for assessment specialists and indeed those interested in feedback on writing in general. As McNamara (1996) says, we should seek information on the scale and the rater with the same rigor as we did for the instrument and subject in the past.

Following McNamara (1996), then, we distinguish the roles of: 1. the 'performance' (in our case essay script) produced by the candidate doing a particular writing task (the 'instrument'); 2. the scale and criteria used; 3. the rater, with all the types of knowledge that he/she brings with them.

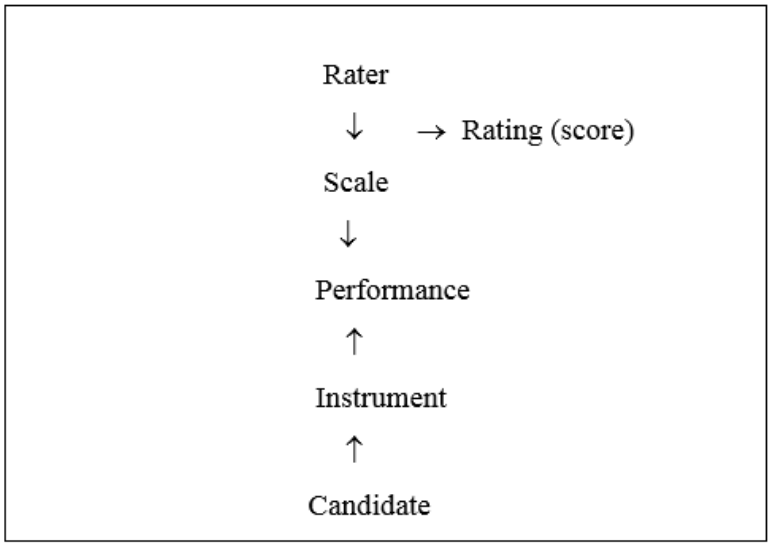

Figure 2-1 Factors in performance assessment (adapted from McNamara, 1996, p. 9)

In more detail (based on McNamara, 1996):

1. There is the target of the rating. This is primarily the text of the composition that is being rated, or as much of it as the rater has read at any given point as he/she performs the rating. This text has qualities or lack of quality in various respects which presumably influence the score or rating assigned, and maybe the process of arriving at it. These qualities are of course in turn the product of characteristics of the student writer, such as his/her language proficiency, background knowledge (schema) of the topic of writing, and the genre (relative to the level and type of the assigned writing task), and writing strategic competence, together with performance features of the act of writing such as concentration or fatigue the writer experienced while writing and so on. We must note, however, that the rater, when rating a given text, may be influenced not only by the text being rated at the time, but also by texts previously rated on the same or other occasions, and what ratings he/she assigned to those, and any other background knowledge possessed about the writer, if known to him (see 3 below).

2. There are the conditions and rules for performing the rating task, at least when performed in exam conditions. These again consist of a number of elements, almost all of which would in exam conditions, and some non-exam conditions, be imposed on the rater from above (an examining board, a school etc.). However, the nature of our study is such that almost all of these are unspecified so are left to the rater (factor 3e) or the rating process itself, rather than independent factors that affect or constrain that process as in most studies. These factors include notably:

a. the purpose of the rating and the 'stakes' associated with it. E.g. for a high stakes exam, a mid term test, or just a practice assignment where the rating may be optional since the focus is on other feedback
b. the audience of the rating. E.g. rating may be done primarily for information of the student, or the institution, or for parents, future employers etc.
c. the overall instructions or rubric for the writing task and for performing the rating based on that
d. the scale or scales (if analytic) on which ratings or scores have to be assigned, whether numbers, letters, or just adjectives like *excellent*, whether holistic, analytic, primary trait etc.
e. the criteria to be used in assigning ratings, and any weightings associated with those. E,g, organisation may be given more weight than mechanics in arriving at an overall score
f. the procedure to be followed when reading scripts and assigning ratings. E.g. it may be specified that content should be looked at first.

In professional examining of writing, and many schools etc., all those are specified except perhaps the last, so are constraints on the rater, who has to interpret and apply them. At the other extreme, in a study such as ours, almost all were left to the raters,

except perhaps a and b, since our study was clearly targeting low stakes rating done primarily for the students' benefit on practice compositions. We are particularly interested in what our participants do with respect to d e f above.

3. There is the agent of the rating, the rater him/herself. He/she has many characteristics which may affect the process and product of the rating, such as:

a. their own writing proficiency, including language proficiency, knowledge of topics and genres, familiarity with different writing task types, etc.
b. L1 cultural and linguistic writing/rating related factors, given that the rater is not necessarily a native speaker of English in which the compositions are written
c. their prior experience of rating compositions (what types and for how many years) or having their own compositions rated, and general professional background as a teacher
d. their training, if any, in rating writing, including what particular rating systems or strategies they may have been introduced to
e. their personal beliefs about what the criteria are for good writing, what is a suitable scale to use, and how to carry out the process of reading and rating to arrive at scores etc.
f. individual personality factors, including inclination to be easy going or strict
g. their knowledge of, or beliefs about, the writers of the compositions being rated, apart from what is apparent in the scripts themselves. This may be specific knowledge of individuals or general perceptions of whole classes or nationalities of students, their needs and wants etc.
h. performance factors affecting the rater, such as boredom or fatigue. Notably these might change over the period of rating a number of scripts.

In our context, we are especially interested in c d e g, as reflected in our research questions (see end of chapter).

Factors such as those listed above are widely recognised in the literature. Thus, Wolfe, Kao, & Ranney (1998, p. 469) recognise that differences in raters' judgments stem from three likely sources: how raters' understand and interpret scoring criteria (our factor 2), their interpretation of the text they assess (our factor 1), and their differences in the rating process (our factor 3).

Huang (2010) highlights how the different linguistic and cultural backgrounds of ESL students make assessment of their English writing a problematic area. On the one hand, many factors affect ESL students' writing, including, their English proficiency, mother tongue, home culture, and style of written communication (factor 1). On the other hand, however, in rating ESL students' written text, raters may differently consider these factors (factor 3 e-g). Empirical studies have found that raters' linguistic backgrounds, previous experience and prior training in the assessment (factors 3 a-d) are amongst factors that affect their rating of an ESL student's writing. The influence of these factors may lead one to question the accuracy, precision and eventually the fairness of the exam assessment of ESL students' writing.

In this vein, Barkaoui (2010) also points out in his study that it may be useful to perceive the rating process as involving a reader/rater interacting with what he sees as three texts 'the writing task, the essay, and the rating scale' within a specific sociocultural context (e.g., institution) which often specifies the criteria, purpose and possibly process of reading and interpreting the three texts to arrive at a rating decision (Lumley 2005; Weigle 2002). However, various claims and many suggestions have been advanced regarding other factors that have a crucial influence on rating and can contribute to

variability in scores and the rater decision-making process. Nevertheless, they can usually be analysed in terms of our three factors above as tending to focus on task requirement (our factor 2), rater characteristics (factor 3) and/or essay features (our factor 1) (Barkaoui 2007).

2.3.2 The need to understand the factors affecting the rating process

The area where the need to understand the rating process has traditionally most prominently arisen and been researched is in connection with testing/examining. A central concern of testers is with reliability and validity of measurement. Essentially reliability is achieved if an essay is always judged to be of the same quality regardless of who rates it or the precise rating system used. Validity is achieved if the rating awarded can be shown to be a genuine reflection of the true quality of the essay.

With respect to reliability, in performance assessment a candidate is assigned a task. This task requires certain behavior to be evaluated on a scale in the light of given rating criteria. As Alderson, Clapham and Wall (1995, p.128) emphasize, a candidate's score on a test must not be subject to who marked the test, or the consistency of an individual marker. That is, ratings must be reliable. The reliability, for example, of a writing assessment is however often 'affected by variations in the perceptions and attitudes of those who read the essays, and the kind of training they receive for reading writing assessment' (Hamp-Lyons, 1991a, p.8), i.e. aspects of our third factor above.

Two types of reliability are often measured in order to characterize raters' behavior. Inter-rater reliability indicates the consistency among different raters in their judgment of candidates' performance (Davies et al., 1999). The other type, called intra-rater reliability, can be defined as 'the extent to which a particular rater is consistent in using

a proficiency scale' (ibid, p. 91).

Reports on the deficiency of reliability of rating in subjectively scored tests have been produced for many years. It is in the nature of open response tests assessed by raters that there will be variation in the score given to different test takers by different raters in response to different test tasks. While it is to be expected, and desired, that different scores would be awarded to different test takers and different writing tasks, it is not desirable that scores should vary dependent on other factors such as the scale used or the rater/tester. However, numerous studies of rater behavior in performance assessment tests have pointed to considerable degrees of undesirable rater variability – variability that is related to some characteristics of the raters (our factor 3) and not related to the performance of test takers (factor 1) (Engelhard, 1994; Bachman *et al*., 1995; Lumley & McNamara, 1995; Weigle, 1998; Congdon & McQueen, 2000; Engelhard & Myford, 2003; Eckes, 2005b; Lumley 2005; Schoonen, 2005).

As we shall see later, training is not always effective. Clearly, much still remains unclear about what raters actually do when they assess testee performance on tasks such as writing tasks.

In our study, we will also be concerned with seeking evidence of reliability and validity of rating, but in different ways from testing researchers. Properly conducted tests/exams always involve all raters using the same rating scale and criteria with which they are supplied, so in order to assess reliability it makes sense to compare raters for how far they assign the same scores to each testee (inter-rater reliability). In our case, however, each rater will be using a different scale and criteria of their own choice or devising, so we cannot assess reliability in that way. We could, however, require raters to themselves re-mark each student/script after a time interval, without looking at their

previous rating, and see how far they were internally consistent by awarding the same rating both times (intra-rater reliability). For practical reasons this was not possible to organise, however, so reliability of rating in our study is not addressed by such standard methods found in testing research, but in a looser sense, e.g. by considering if a rater tends to tackle the rating process of each script in a similar way or not. For similar reasons, while our study will consider effects of prior training on rating, it will not be doing so in the way testing researchers do, seeking to ascertain how far training on the one rating/scoring system to be used impacts on inter-rater variation in scoring the same testee.

Rater variability or error, or inconsistency between raters, can manifest itself in various forms during the actual process of rating scripts (Engelhard 1994; Bachman & Palmer, 1996; McNamara, 1996; Weigle, 2002; Weir, 2005; Lumley, 2005), and not only in scores awarded. For example, as they rate, raters may differ (Eckes, 2008):

1. in the degree to which they follow the scoring rubric
2. in the technique, they use to interpret and apply the criteria during scoring sessions
3. in the degree of rater severity or leniency of the rating/scoring, where one rater consistently rates either higher or lower than the quality of the performance deserves, or than other raters
4. in the understanding and use of rating scale categories
5. in the level to which rating is done consistently across test takers, scoring criteria, and performance tasks
6. in randomness or inconsistency in rating behaviour
7. in central tendency, when raters tend to give ratings clustered closely around the mid-point on the scale.

This list again largely presupposes a context where all raters are supplied with the same scale, criteria, rubric etc. Since that will not be the case in our study we will be more concerned with variability/unreliability of types not based on that, such as points 5 and 6 and 7.

We now present a detailed review of some of the influences on rating behaviour which are most relevant to the present study, related to the rater, and to the rating task (i.e. aspects of factors 2 and 3 in 2.3.1). After speaking with the course coordinators and exploring the two institutions where the study is located (see details in Chapter 3), these factors seem likely to play an important role in the rating process and decision making in the context of our study. Hence, it is worth giving them space in this chapter so as to be able to refer to this literature when discussing our results.

2.3.3 The nature of the rating scale and criteria as a factor in rating

We first review some elements of how the rating task requirements (our factor 2) impact on the rating process. As we have indicated, often, and certainly in formal tests and exams, they are imposed on raters, and most of the literature assumes that; thus, we have not been able to find extensive discussion of teacher self-chosen or informal rating, which might not even use a numerical scale. In our study we anticipate, however, that raters may spontaneously adopt some elements of what we discuss here if they believe in them, perhaps due to previous experience or training.

2.3.3.1 Definition of a rating scale

A rating scale is scale with set of descriptors associated with bands or points (indicated by numbers or letters) to define the proficiency level or language ability associated with each. Such scales play an important role in what is termed performance assessment,

including assessment of writing, where response is open rather than fixed choice. They are used to enable raters to make judgments on candidates' performance according to them and also, they help to define levels of language ability for the benefit of the testee. The former is the "assessor-oriented" guiding purpose whereas the latter is called the "user-oriented" or reporting purpose (Bukta, 2007, p.71). In our study, where the rating is not done as part of a test and the scale is not externally imposed, the second purpose is more relevant. The "assessor-oriented" purpose is more relevant to rating done as part of testing and is supposed to exhibit reliability of rating, make sure the rating is standard, and help to establish a common ground for score interpretation by raters. Hence, the rating scale can help others to make inferences about the language ability of candidates.

2.3.3.2 Alternatives to rating scales

Hamp-Lyons (1987) mentions that Cooper (1977) divides scoring procedures for direct essay assessment into two broad categories which he called 'holistic' and 'frequency count'. She quotes that he describes 'holistic' methods of evaluation as any procedures which "stop short of enumerating linguistic, rhetorical, or informational features of a piece of writing" (cited in Hamp-Lyons, 1987. p.96). This therefore includes rating of both the types distinguished today as holistic and analytic. For Cooper, 'frequency count' methods include 'error counts', 'T-unit counts' and other methods in which features of the writing are enumerated. As Cooper admits, there is some difficulty in distinguishing the most atomistic of the 'holistic' methods (in today's terms, highly analytic ones) from the most integrative of the 'frequency count' methods.

Frequency counting methods of assessing writing have rarely been used because of the time and labour involved to count relevant linguistic elements, which makes them

impractical both for professional exam markers and classroom teachers dealing with practice assignments. Interestingly, however, assessment of writing through counting features of text is currently enjoying a revival as researchers attempt to develop convincing computerized systems for scoring writing by automatically counting features of it (Callear et al., 2001).

Furthermore, sometimes the definition of a point on a rating scale in effect assumes that the rater does a rough count of some specific feature, e.g. where a B for mechanics is defined as 'the writer makes very few spelling errors'. In our study, we will be interested to see if our teachers, unprompted, spontaneously quantify any aspect of the essays in their process of rating.

2.3.3.3 Types of rating scale: overview

Mullis (1984) categorizes methods of scoring direct writing assessments into three methods, holistic scoring, primary trait scoring, and analytic scoring, as the most three frequently used types of scale in written performance assessment at least by one rater. She appears to discount 'frequency count' methods altogether. The choice among them depends on the preference whether we want to choose to assign a single score or more scores (holistic versus analytic), and whether we want a scale that in principle applies to all aspects of all written texts (generic), or one that is tailored to rating only a certain aspect that may be of interest for a particular writing task given at a particular time to particular students (primary trait). For example, if a teacher has just been teaching connecters, and sets a writing task, she may rate the next compositions written just for their use of connecters. Weigle (2002) believed that every single writing task can possibly have a specific scale, or the same scale can be used for assessing different tasks.

Primary trait rating has been extended to multiple trait rating (Hamp-Lyons, 1991), where more than one selected feature is rated, but both share the characteristic of being task specific in comparison with what may be called generic rating. We now look at each of these separately.

2.3.3.3.1 Holistic scales

Most researchers into performance assessment today make a primary distinction between holistic and analytic scales (for example Alderson et al., 1995). According to him, a holistic scale is used to assess performance as a whole, whereas the analytic scale provides the rater with the opportunity to look at the elements of performance and assess them separately.

Hamp-Lyons (1987) defines holistic scoring as "[involving] reading the essay for an overall impression of the quality of the writing and assigning it a score based on this overall or 'global' quality" (p.96). In this process, raters assign scores matching the scripts to a scale with several proficiency levels and each level is described. Moreover, Hamp-Lyons (ibid) believes that holistic scoring serves to reflect the quality of any piece of writing more than the quality of any of its directly observable parts, and that undefined quality is something that skilled readers can recognize.

Hamp-Lyons (1992, p.2) provided a specification of how writing assessment using a holistic scale should be conducted for exam purposes, with five components. Initially, in order to assess written performance holistically, candidates ought to produce one or more texts of at least 100 word each. Then, although the writing task rubric defines the task and provides some kind of prompt, candidates should have the right to compose the texts in the way they like. Thirdly, rating is realized using one or more raters. Fourthly, rating is based on agreement between raters or on sample scripts or on a rating

scale. Lastly, raters' decisions are expressed in scores, which can be later interpreted to make inferences about candidates' language ability. Normally, holistic raters as readers are asked to read rapidly, creating a global judgment and not focusing on any specific features such as organization, mechanics or ideas.

Like any other methods, holistic scoring has several advantages: it has been claimed to have the highest construct validity when overall attained writing proficiency is the construct to be assessed (Perkins, 1983). The performance is evaluated as a whole; its overall effect on the rater is at the center. In this method, scoring is faster as there is only one scale to attend to. Additionally, it used to be regarded as frequently used as a tool for "certification, placement, proficiency, and research testing" (Perkins, 1983, p.653). In this last respect, Perkins is surely out of date, however, since standard international proficiency tests today, such as Cambridge (including IELTS), all use some form of analytic scale.

It is generally acknowledged that with careful rater training and monitoring, this kind of scoring procedure can produce reliable results (McNamara, 1996; Weigle, 1994, 2002). However, these rating processes have been criticized for oversimplifying the constructs they are supposed to represent. As Cumming, Kantor, and Powers (2002) explained,

> "Holistic rating scales can conflate many of the complex traits and variables that human judges of students' written composition perceive (such as fine points of discourse coherence, grammar, lexical usage, or presentation of ideas) into a few simple scale points, rendering the meaning or significance of the judges' assessments in a form that many feel is either superficial or difficult to interpret." (p. 68)

Further, this way of rating arguably does not provide sufficient information on the components of language ability as a single score cannot reflect all aspects and it "[reduces] the writers' cognitively and linguistically complex responses to a single score" (Hamp-Lyons 1991, p.244). Therefore, Hamp-Lyons (ibid) admits that whereas there are serious problems with holistic scoring in any context, these problems are specially evidenced in ESL writing assessment contexts. These problems can be summarized as that raters may attend to different features and arrive at different scores which are then the only sources for making inferences about language ability.

Additional to the theoretical difficulties with this method, the nature of holistic judgments presents practical weaknesses, since the scores generated in this way are often not accompanied by any definition or characterisation of each scale point, so cannot be explained either to the readers or to the people affected by the decision made through this scoring process. The danger of this is that this scale hinders the production of reliable scores if other readers of the same assessment community are involved in the rating process. In order to obtain more detailed descriptions of candidates' writing ability, primary and multiple trait scales or analytic scales can be used (Weigle, 2002, p.112-114).

In our study, we will therefore be interested to see whether teachers in a non-examining context, and with no scale provided, in fact favour some kind of generic holistic scale like this, despite its possible ambiguity of interpretation for the student writers, or not.

2.3.3.3.2 Primary and multiple trait scales

A primary trait scale is a specific type of holistic scale which, along with multiple trait scoring (a form of analytic), is mostly used in L1 assessment. Hamp-Lyons (1987) believes that primary trait scoring falls between holistic and analytic scoring. In this

procedure, criteria are clearly defined and levels are specified, as in analytic scoring, but only one judgment is made, as in holistic scoring.

Lloyd-Jones (1977), who drew attention to primary trait scoring, makes it clear that primary trait scoring is based on the idea that one should judge whether a writing sample is good or not by reference to its exact context and that suitable scoring criteria should be developed for each writing task prompt. Hamp-Lyons (1992, p.8) states: "the theory is that every type of writing task draws on different elements of the writer's set of skills, and that tasks can be designed to elicit specific skills". Multiple trait scoring has the same premise but focuses on several different aspects and makes awarding several scores for each script possible, as in analytic scoring.

A primary or multiple trait scale then focuses on key features of each script only, and is task specific (Cumming et al., 2002, p. 68). Students are assigned to write, for example a persuasive essay and are assessed on the degree of fulfilment of that task only. This method of scoring is designed separately for each writing task and involves the writing task, the rhetorical features, expected performance description, a rating scale, sample papers and explanations of scores on sample papers (Hamp-Lyons, 1992). Although this method of writing performance assessment would be beneficial in L2 contexts, it is not widely used (Weigle, 2002, p. 110-112). Possibly the amount of work involved in creating a different scale, criteria etc. with definitions and so forth for each writing task could be the reason.

We will be interested to see in our study if teachers, without being provided with any scale, and rating their own students' work, in effect pick on one or a few criteria to base their rating on, maybe related to the specific occasion, topic or task type of the writing assignment they set, or adopt some form of generic rating (whether holistic or analytic)

which aims to be applicable regardless of task, occasion etc. Although primary and multiple trait scales are not widely used for formal tests and exams in ESL contexts, it is after all possible that teachers for their own classroom feedback on practice assignments in fact do make use of it.

2.3.3.3.3 Analytic scales

Bachman and Palmer (1996), referring to generic analytic scales, state that an analytic scale "requires the rater to provide separate rating for the different components of language ability in the construct definition" (p.211). An analytic scale gives more detailed information on candidates' performance than any type of holistic scale defined above. The writing performance in this scale is divided into several aspects, levels, or criteria, for instance, content, vocabulary, accuracy, organization, etc., each of which is assigned descriptors according to proficiency or achievement levels. These levels can have the same or different weighting in the calculation of an overall score from the individual ones, which is determined by the importance assigned by whoever devised the scale to the aspect in question. A candidate may therefore get higher or lower scores on different aspects reflecting the differences between the components of language ability (Alderson et al., 1995, p.108).

Palmer and Kimball (nd, cited in Hamp-Lyons, 1987, p.101) developed an analytic 'Criterion-based Composition Grading System', consisting of nine characteristics each described on a very simple scale. The 'Composition Profile' of Jacobs et al (1981), uses a fully described analytic scale with components differentially weighted. Developed for use in scoring college-level ESL writing, the Profile has been carefully worked out and extensively validated and widely used in writing research. It includes five components or types of criteria (content, organization, vocabulary, language use (=grammar), and mechanics) each of which is rated at a number of levels and a brief indicator of the

characteristics of writing at each level in each section is provided. The scores for subscales can be either combined or reported separately. Weigle (2002, p.114-121) highlights the importance of explicitness of scale descriptors and that distinctions between levels should be clear. As a result, analytic scores reported show an informative picture of test takers' language ability. On the other hand, Weigle (ibid) once again highly recommended rater training regardless of the selected scale, as training in using the rating scales hopefully ensures reliability of judgments.

Cumming and Riazi (2000) also surveyed indicators of achievement in writing in L2 and they selected an analytic rating scale as a measurement tool. They found that the elements of language ability were shown in a more detailed way than by any other assessment instrument. They further claim that an analytic rating scheme is multi-faceted similarly to the ability it intends to measure.

2.3.3.3.4 Holistic vs analytic scoring

Both holistic and analytic types of rating have gained wide acceptance in writing assessment in large scale and classroom assessment (Hamp-Lyons, 1991; Weigle, 2002). As we have seen, however, these two types of scales vary in terms of scoring methods and implications for the rater decision-making process (Goulden, 1992, 1994; Weigle, 2002). Barkaoui (2010) speculates that these differences are likely to influence the essay rating process and outcomes. However, the literature is replete with arguments for the advantages and the disadvantages of both rating methods.

The dilemma of which rating scale to apply for written performance assessment has indeed been intriguing researchers for some time now. As far as reliability and validity of assessment are concerned, the scope of choice is usually limited to either analytic or holistic scales, omitting primary trait or frequency methods, as both the former are seen

to be most appropriate for measuring written performance (Shaw, 2002). The common feature of all analytic scales is the opportunity to focus separately on the qualities of the writing which are important for the purpose of the assessment. This analytic endeavour at precision is in contrast to the imprecision which is counted as a virtue by the proponents of holistic scoring. Brown (1981) says:

> "No matter how reliable holistic scoring is as a way of rank-ordering papers, it is inadequate as a measuring tool in itself, because it is relativistic and is not tied to any absolute definition of quality" (cited in Hamp-Lyons, 1987, p.102).

We would argue that this is a mistake, however, since it is perfectly possible for holistic scales to be accompanied by detailed, if very broad, descriptions of what each point on the scale corresponds to in essay quality, analogous to the definitions associated with points on analytic scales.

A comparison of the two types of rating scale on six qualities of test usefulness was proposed by Weigle (2002, p. 120). The following qualities were utilized by Weigle (2002) to compare the two scoring scales: *reliability*, *validity, practicality*, *impact*, *authenticit*y and *interactiveness* (Bachman & Palmer, 1996, p.17-43).

Bukta (2007) reports a detailed summary of the findings as follows. *Reliability* of the holistic scale is lower than of an analytic one. By contrast, holistic scoring has the highest construct *validity* when overall attained writing proficiency is the construct to be assessed (Huang, 2010). On the other hand, it has "threats to reliability" because it can be highly subjective due to "bias, fatigue, internal lack of consistency, previous knowledge of the student, and/or shifting standards from one paper to the next" (Perkins, 1983, p. 653). In short, there is no doubt of the important role of the scoring rubric and a rating scale accompanied by definitions of scale points in writing

assessment as they contribute to higher reliability (Conner-Linton 1996; DeRemer 1998). As we have noted, however, reliability may not be uppermost in our raters' minds, since they are not rating for test/exam purposes.

Additionally, a holistic scale is more *practical*, as scoring is fast and easy, which could make it attractive to the teachers in our study who are rating practice assignments as part of feedback to students, rather than exam compositions. Analytic rating is more time consuming but it produces higher inter-rater *reliability* than holistic scoring (Perkins, 1983; Bukta, 2007; Huang, 2010).

Regarding *impact* issues, this includes score interpretations. Holistic scores usually provide less information on writing ability and decisions are more difficult to be made. Possibly, therefore, our teachers may prefer some kind of analytic scale because it gives more detail, so is more relevant where feedback to students is the main aim. Holistic scoring has higher *authenticity*, however, as reading scripts in this method of scoring resembles real-life reading more than reading scripts for analytic scoring.

Lastly, the two scales were not compared for *interactiveness* by Weigle (2002), as interactiveness relates to the relationship between the test and the test taker, so it is not applicable for rating selection for formal assessment where the testees do not usually receive detailed information on how their scores were arrived at (Weigle, 2002, p.121). We should note, however, that in the kind of rating that will be considered in our study, which is not done for formal exam assessment but as part of formative feedback to learners during a course, this feature could be important. We will therefore look to see which type of scale teachers in fact choose in this context.

Empirical evidence supports the views stated above that the rating method may affect the reliability and validity of the rating of ESL compositions. In this regard, Song &

Caruso (1996) conducted a study which examined the degree to which differences existed between holistic rating and analytic rating of four compositions written by two Russian ESL speakers and two NE speakers. It was found that the holistic and analytical methods produced no significant differences between the scores assigned to ESL and NE essays. Nevertheless, the holistic scores given by the English faculty and those assigned by the ESL faculty differed significantly with the English faculty assigning higher scores to all four essays. The English faculty raters seemed to give more weight to the overall content and quality of the rhetorical features in the writing samples than they did to language use.

Additionally, other studies support the above differences between the two scales. Nakamura (2002) asked three trained raters to rate 90 scripts both holistically and analytically. The results of both methods were compared bearing in mind the theoretical framework of the differences discussed above. Findings showed that holistic scoring seems to be less costly when it comes to economic reasons. This research concluded that many raters as well as multiple scales are the best practice however. One rater and an impressionistic holistic scale are the least desirable choice. The second best practice this study by Nakamura (2002) concludes is to use holistic evaluation with more raters. This study warns that holistic rating may not to measure level of proficiency accurately and can lead to misinterpretations of student ability. This study maintains the importance of reliability and construct and content validity in making a choice for exam assessment purposes.

In conclusion, we may however question whether much of the discussion in the literature, or the criteria for what is the most suitable scale valued by these studies, concerned with exam assessment, where accuracy is paramount, really correspond with the criteria suitable for classroom teachers doing day to day rating of practice writing

assignments. They often have little time and lack the luxury of a second marker being available; furthermore, they often need a scale which provides useful information to the students rather than extreme accuracy of score, and may on occasion not require generic rating at all, but find primary or multiple trait rating more appropriate, to suit the current focus of instruction. In any event, we look forward to seeing what the teachers in our study favour in these respects.

2.3.4 The nature of the rater as a factor in rating

The effect of rater background (our factor 3 in 2.3.1) on the rating process is central to some of our research questions and our participants will be selected to have different backgrounds, as found in the study context. Hence, it is felt that it is crucial to discuss this issue.

Moreover, the significance of aspects of raters' background has been clearly evident in many studies investigating the rating process (e.g. Cumming 1990; Weigle 1994; Cumming et al. 2001; Erdosy 2000; Zhang 1998). Erdosy (2000) for example argued that some studies of language assessment and rater behaviour have shown that culture, mother tongue, academic background, and professional experience have a strong influence on rating behaviour. Gender, professional and experiential backgrounds, the amount of training in using assessment tools and the amount of exposure to L2 writing are also cited (Hamp-Lyons, 1990; Vann, Lorenz and Meyer, 1991). These factors will now be discussed in detail.

2.3.4.1 Rater professional background

A series of essentially qualitative studies has identified differences in rating behaviours as a result of professional background (O'Loughlin 1992; Elder 1993; Deville 1995,

1996). In contrast, Hamp-Lyons (1989, p. 239) found that "we cannot even say with certainty that ESL and non-ESL raters are valuing something differently, whatever it may be", while a study of a 'specific-purpose' tests of speaking (Lumley, Lynch & McNamara 1994; Lumley 1998) showed no difference in rating standards applied by groups of ESL-trained assessors and medical practitioners.

2.3.4.2 Rater cultural background

Cultural background is amongst factors that influence ESL compositions' raters in addition to mother tongue and professional experience. These factors appeared in studies to exert particularly strong influences on rating behaviour, yet, however, none could explain it in isolation from other factors. Since in our study cultural background will be largely constant we do not pursue this aspect further.

2.3.4.3 Rater professional experience of rating

The writing assessment literature has shown that raters' professional experience such as number of years of teaching and rating ESL written texts influences their rating (Cumming, 1990a; Hamp-Lyon, 1996; Rubin & William-James, 1997; Vaughan, 1991). These studies have shown that the professional experience emerges as an important variable in two ways.

Firstly, raters may be exposed to very different learner populations as teachers, guiding them to form different anticipations from learners. This is easily demonstrated through documented differences between the rating behaviors of teachers of English as a first language and the rating behaviors of teachers of English as a second language (Erdosy, 2004). Teachers of English as a first language have been discovered to be consistently more severe than ESL teachers in their judgment of sentence-level errors, which may

be the only generalization with widespread empirical support in the existing literature.

Secondly, in terms of years of experience, Song and Caruso (1996) found that raters with more experience of teaching tended to be less strict than raters with less experience of teaching when they used holistic scoring. Contrarily, years of experience did not significantly affect their rating of ESL compositions when using analytical rating scales.

Weigle (1994) further demonstrates that scores awarded by inexperienced raters are affected by rater training, such that inter-rater consistency is improved. Hamp-Lyons (1990, p.81) suggests that studying how the "experiential backgrounds" of raters may influence their responses has suffered as a consequence of a major preoccupation with testing and validation of scoring procedures. Partly for this reason we are including raters with varied experience, and training, in our study.

2.3.4.4 Rater linguistic background

Raters' linguistic background (native language) can also be a factor that affects their ratings of ESL compositions. The influence of such a factor emerges most clearly in the contrasting attitudes of native speaker and non-native speakers to ESL compositions, and the growing number of NNS assessors, yet this difference is important (Erdosy, 2004).

Kobayashi's study (1992) examined 145 native English raters and 124 native Japanese raters at the professorial, graduate, and undergraduate levels rating two compositions written by ESL students at an American university. The results should that NE raters were stricter with respect to grammar than the native Japanese raters, and that NE professors and graduate students rated more positively clarity of meaning and

organization in both compositions than the native Japanese raters did. On the other hand, the native Japanese undergraduates rated both compositions much more positively than did the NE undergraduates

In our study, all but one of our raters are NS of English, so this factor is less in focus.

2.3.4.5 Rater reading/rating style

Another factor that has an influence on the scores or decision-making process in second language assessment is recognised to be their rating or reading style (Cumming 1990; Vaughan 1991; Milanovic & Saville; Milanovic, Saville & Shen 1996; Smith 1998, 2000; Cumming et al. 2001). Rating style here refers to the often individual process by which a rater typically reads the essay, interprets the rating scale, where one is provided, and assigns a score (Lumley 2005; Sakyi 2003; Smith 2000, cited in Barkaoui 2010).

Lumley (2005) has pointed out that it is important to recognise the relationship between reading the text, interpreting the scale and awarding scores in order to understand the rating style. The importance of identifying the rating style or reading behaviour of raters emerged from the work of Sakyi (2000), who observed four different reading styles in action when individual raters evaluated essays.

Hamp-Lyons (1990, p.81) more broadly suggests that the background of a rater is not just a matter of demographics and professional background, but also a whole range of aspects relating to the processes of rating including the decisions raters make, how raters use rating tools, the experience of reading, etc. and how they affect rater judgements.

It is worth emphasizing that rating style, in the sense of the process the rater habitually follows when rating a script, can be seen as a background characteristic of the rater, as

can their rating strategic competence and indeed their preferred scale and criteria (next subsection). However, insofar as these have been studied, it has been usually not as a background feature of the rater (part of their beliefs or competence) accessed through questionnaires or general interviews. Rather these have been studied 'in action' (as part of the rater's practices or performance) when the rater is using his/her background characteristics while actually performing a rating of a script. This is often accessed through think aloud data. Hence this will be taken up again in 2.3 (especially 2.3.1.3).

2.3.4.6 Rater preferred scale and criteria

Similar to having a habitual reading style used when rating, raters may have their own beliefs about what criteria are relevant to rating writing and their relative importance/weighting. This may be voiced in terms of their idea of 'what makes a good essay' or the like. Such ideas will interact with any criteria provided when producing a rating for a particular script during the rating process, or if, as in our case, no criteria are provided, then they will have free rein. There is rather little literature on this, which, as indicated in chapter 1, prompted us to include this in our study.

2.3.4.7 Rater training

Undoubtedly, rater training is a fundamental element in the essay rating process (Davidson, 1991; Weigle, 1994). Training provides raters with a clear conception of what a piece of quality writing looks like, within the criteria of some specific rating/scoring system, and accordingly stimulates rater consensus (Homburg, 1984). Additionally, Jacobs et al., (1981) and Reid & O'Brien (1981) believed that training can minimize differences in scores given to compositions caused by raters' different backgrounds as indicated above and adjust expectations of good writing by clarifying for the raters both the task demands and writer characteristics (Huot, 1990). As a matter of fact, as we noted earlier in 2.3.3, rater training is an issue which lies at the heart of both reliability and validity in ESL essay rating (Weigle, 1994). Homburg (1984) commented that holistic rating of ESL compositions, "with training to familiarize readers with the types of features present in ESL compositions, can be considered to be adequately reliable and valid" (p. 103).

There are, however, some researchers who question rater training as a necessity, as in Purves's (1992), words "No matter how extensive or thorough it may be, the rating is still a perception, a subjective estimate of quality" (p.118). Moreover, rater training has been revealed in some studies to be much less effective in lessening rater variability than expected; that is, raters habitually remain far from functioning interchangeably even after extensive training sessions (Lumley & McNamara, 1995; Weigle, 1998; 1999; Hoyt & Kerns, 1999; Barrett, 2001) or after individualized feedback on their ratings (Elder *et al*., 2005). Song and Caruso (1996) also found that training did not differentially affect the way raters were using a holistic scale, compared with an analytic one, while number of years of teaching experience did have an impact.

Although, this account reveals the complexity of the factors affecting the rating process, yet the value of rater training cannot be neglected, certainly for testing purposes. Hence, we will be looking in our study not only at what relevant training the participants had received but also their views on training.

2.4 The rating process

Having discussed key factors which can affect the rating process, and the ratings/ scores that emerge from it, we now turn to research on the nature of the process itself, which is targeted by several of our research questions. Here the impact of those various factors may be seen more clearly as a rater proceeds through the real time activity of reading a script, applying criteria, and arriving at a rating on whatever scale is being used, influenced potentially by a range of his/her background characteristics.

Over recent years, L2 writing researchers have consistently recommended that investigating the processes raters go through in arriving at judgements is one way to reach a greater understanding of rater behaviour (Huot, 1990, 1990a; Hamp-Lyons, 1990; Tedick and Mathison, 1995; Brown, 1995 and Milanovic, Saville and Shuhong, 1996). Therefore, in the sections below we review some key studies. In order to gain a deeper insight into these processes, qualitative data has typically been collected using some kind of introspection, often via think aloud reporting, while quantitative data may take the form of scores awarded by raters to essays, or counts of how often various strategies were used during the rating, by the same person or different ones.

It should be noted that studies vary considerably in how they talk about the rating process, however. For instance, some, such as Wolfe, refer to it as rater cognition. Also, the individual activities that the rater engages in during the process are variously termed comments, behaviours, processes, or strategies. Once again, we must also bear in mind

that much of the research had in mind the rating of essays as if for examination purposes, as distinct from the scenario of our study. Often these studies are also therefore interested in the product of the rating process, i.e. the ratings or scores awarded, which are of less concern to our study.

Many issues have come to the fore as of interest in relation to the scoring or rating process itself. Cohen for instance points out that one main line of research in the area of rater behaviour focuses on the different degree of raters' attention to the various features of text, such as content, mechanics or organization, and their scale interpretation (Cohen, 1994b, p. 332-336). Shaw lists the following as fundamental questions (Shaw, 2001, p.3):

- What is the raters' decision-making behaviour in terms of the approaches they employ while marking EFL compositions?
- What elements do the markers focus on while marking compositions?
- How does examiner marking behaviour compare in relation to inter-rater consistency?
- Do examiners adjust their marking behaviour according to the level of the script?

In our review, we will not be able to deal with all the many issues that have been discussed, and in any case not all are relevant to our study (e.g. inter-rater consistency). Rather we will focus on areas that have implications for us, which for convenience we divide into issues concerning strategies used in the rating process (2.3.1) and issues concerning sequencing of strategies during rating (2.3.2).

2.4.1 Key empirical studies of the strategies in the process of rating English writing

In this section I look at key previous studies' investigations of the strategies or behaviors that are evidenced in studies of the rating process, as distinct from just looking at the scores that emerge from that process, often with attention also to the impact of selected factors from those we covered in 2.2.

This section will consider studies both in English mother tongue context assessment and assessment of second language writing.

2.4.1.1 Huot (1988)

Huot (1988) and Pula and Huot (1993) conducted ground breaking and much cited work which investigated rater behaviour in L1 English writing assessment. In this work Huot looked at trained and untrained raters' rating and also, importantly for us, rater behaviour. Huot opened up the topic of the natural behaviour of raters and investigated it in detail in the context of a holistic rating session (Lumley 2005). This investigation was carried out using think-aloud protocols, which is one of the early methods that was used in investigating the rating process. In Huot's study, he declared that "the personal nature of reading and evaluating cannot be ignored; it is part of the process of making meaning from text and interpreting that text to make a judgment about its worth." (1988, p.27).

Huot (1988) hypothesized that the responses to a written text evaluated by naïve raters applying their own criteria will vary and exceed in number those provided by trained raters using criteria imposed on them in a scoring rubric, albeit they had composed this rubric collaboratively for themselves. Thus, his study is not a simple comparison of trained and untrained raters but also of self-chosen and imposed criteria. Still the

untrained raters were in a similar position to the raters in our study. We will however improve on his design by comparing both trained and untrained raters in a situation where neither is provided with any scale or criteria to follow.

Surprisingly, the result of his study contradicted his expectations: Huot (ibid) interestingly found that the trained and untrained groups adopted essentially the same criteria for their ratings. These represent findings which we may or may not replicate in our English L2 writing situation. Huot (1988) further claimed about the novice raters that their rating strategies derived from their reading process while rating the essays, which again is something we will look to find also. In his study Huot (1988) also claimed that he shared findings with other studies (Charney, 1984) which counter any negative view of holistic scoring procedures as being a hindrance to a true and accurate rating. On the contrary, in his view holistic scoring might actually yield a rating process that guarantees a valid reading and rating of student writing.

One implication that can be drawn from this study in Huot's eyes seems to be that the scoring rubric may be somewhat redundant, exercising little influence upon the rating process in holistic rating, since both groups essentially used the same criteria. Verification of this finding was made in a partial replication by Pula & Huot (1993) "…. since both trained, experienced raters and English teachers without prior training or experience in holistic scoring procedures based their rating decisions on very similar criteria". (p.237). The implication drawn is that the experience of teaching common to both groups of raters was the source of the criteria employed. This then presents a possible expectation for our study where we have no rubric or criteria provided and indeed not even a rating scale, and a mixture of teachers with different experience and training, but who all teach on the same kind of writing course. Will they too in fact be found to use the same criteria?

Beside illuminating the evaluative process, Huot's study was one of the first to document that the rating process contains some comments which are not evaluative, where raters observe and comment on features that he claims do not contribute to the score they give. He appears however to see only judgmental comments that are explicitly recognized within the evaluation criteria, as relevant to the score. According to Huot (1988), 'personal comments' instead epitomize 'engagement', a process more like natural reading. Hence these comments have a certain validity, although he eliminated anything he labeled as 'personal comment' from consideration as an influence on perceptions of quality as represented by the score given. This area is one we feel to be important, so we will be formulating a research question specifically about such 'other comments' made by raters which are not directly evaluative.

Table 2-1 shows how Huot (1988) and Pula & Huot (1993) coded the categories of 'personal comment'. We can see that raters comment on a variety of aspects of the rating process that occur to them, some of which, despite Huot, we would surely see as related to evaluation and some arguably not.

Table 2-1'Personal comment' categories (Huot, 1988; and Pula & Huot, 1993) (adapted from Huot, 1988; and Table 1. Pula & Huot, 1993, p.244)

Personal comment	
1. indecision[1]	Hesitation about score
2. opinion	Rater's feeling - general, unspecified
3. expectations	Of what's to come in essay - predictions and hopes
4. laughs	
5. nonevaluative	Directed towards content or writer's situation
6. sarcasm	
7. justification	Not clearly defined
8. questions own judgment	Related to rubric

[1] The original table has no numbering; I numbered them to make it easier for the reader.

In Table 2-1, we can see that there is a relationship between categories 1 and 8, with 1 focusing on the score and 8 focusing on the wording of the scoring rubric. In addition, 7 'justification' seems to be evaluation-related although listed as a 'personal comment' other than an integral component of the scoring. In fact, several so-called personal comments appear to be evaluation related. Huot therefore, according to Lumley (2005), with whom we agree, offers a very restricted model of the rating process, in which personal engagement is explicitly considered as separate from the rating process, and actually contains some evaluation related comments. Our study will see 1 7 and 8 as part of the central evaluative process.

Huot (ibid) also found that holistically trained raters "contributed a substantial number more personal responses representing many viewpoints [ ] than raters who read the same papers and were not trained" (p. 237). In fact, the experts made 411 comments whereas the novices made 289 comments, although Huot sees this as not a large number, so not a sign of much difference between the two groups. One possible reason for such a distinction is that the novice raters were dealing with two completely unfamiliar tasks: firstly, rating without benefiting from a scoring guide, and secondly thinking aloud to describe the rating process. These two activities might have led them to produce fewer comments which Huot placed in the personal interaction category.

2.4.1.2 Cumming (1990)

An interest in how teachers assess L2 writing is found in the work of Cumming (1990), who compared the decision-making processes of novice and experienced raters of ESL composition, rating compositions by intermediate and advanced proficiency writers who also differed in writing experience. This study therefore incorporated investigation of selected variables from what we called factor 1 and factor 3 in 2.2.1.

In this study, Cumming employed two research designs concurrently: a) examining 12 texts representing students varying in both L1 writing expertise (average and professionally experienced writers) and ESL (intermediate and advanced) proficiency and b) comparing the rating performance of six experts and seven novice raters. The rater groups produced both scores and concurrent think-aloud protocols describing the process of rating the 12 compositions, using an analytic scale imposed by the researcher - a 4-point scale for each of three dimensions: effectiveness of 'language use', 'rhetorical organization', and 'substantive content' (1990, p.34). Since all teachers rated the same compositions, whose authors were unknown to them, using a prescribed scale, the situation resembled the rating of compositions written for a test more than written for practice for an individual teacher (as in our scenario).

Cumming (ibid) found a statistically significant difference between the ratings of novices and experts in relation to the criteria of rhetorical organization and content (in both cases the novices' ratings were higher); there was no such difference, however, in the way the two groups rated language use. In our study, since raters are not rating the same set of scripts, and not using the same criteria, this kind of result will be impossible to detect. We feel, however, that this is a small loss since many other writing assessment studies have made such comparisons while almost none have considered the kind of naturalistic classroom practice rating which we focus on.

With regard to the rating process as revealed by think aloud protocols, of the 28 distinct behaviours that Cumming identified, 20 could be classified under the three categories or dimensions for which raters were required to award scores: substantive content, language use, and rhetorical organization. The remainder Cumming classified as aspects of what he called 'self-control focus' used by the raters. These included two broad strategies. Firstly, 'interpretation strategies' were used in reading the text (e.g.

scan the whole text to obtain initial impression), as well as in making comments or inferences about the writer's condition or how the writing was produced, for instance whether or not it appeared to have been memorized (1990, p.37, 47). However, those strategies were not evaluative. Secondly, comments made under the 'self-control focus' also include a range of 'judgment strategies' used to evaluate the text in ways beyond the given criteria (Cumming, 1990, p. 37) as in Table 2-2

Table 2-2 Judgment behaviours with self-control focus (Cumming 1990, p.37)

1.	**Establish personal response to qualities of items**
2.	Define, assess, revise own criteria & strategies
3.	Read to assess criteria
4.	Compare compositions
5.	Distinguish interactions between categories
6.	Summarize judgments collectively

This set of behaviours evidently embraces a number of items which would fall into Huot's (1988, 1993) 'nonevaluative' category, discussed above. In contrast to Huot, however, Cumming does not claim that these comments do not impact on the score awarded and we also will in our study regard many of them as evaluative.

Significant differences were found among novice and expert raters in relation to some specific strategies: for example, experts paid more attention than novices to analyzing the rhetorical structures of the text. Overall, however, there was no significant difference in the frequencies with which strategies were used by novice and expert raters. Thus, any overall differences in the rating behaviours of the two groups were more qualitative than quantitative.

Cumming does however make a solid case for the impact on thinking processes of the expertise of the teachers trained as assessors of second language writing:

> "Overall, expert teachers appear to have a much fuller mental representation of "the problem" of evaluating student compositions, using a large number of very diverse criteria, self-control strategies, and knowledge sources to read and judge students' texts. Novice teachers tend to evaluate compositions with only a few of these component skills and criteria, using skills which may derive from their general reading abilities or other knowledge they have acquired previously." (1990, p.43)

Cumming's study confirmed Pula & Huot's (1993) findings that novice and expert raters used different methods in the rating task. In particular, he discovered that the novices often did not make their criteria explicit, and tended to rely for example on judgments of the situation in which the texts were produced, which apparently gave them insufficient grounds for assessing language use, which should rather be based on the texts themselves. In other words, the novices lacked a coherent and principled basis for conducting the business of rating texts. The experts, in contrast, noticed errors in the scripts and classified them to inform their overall judgments about the writing; they did so, however whilst retaining a focus on the content and organization of the text.

> "Expert teachers, on the other hand, integrated their interpretations and judgments of situational and textual features of the compositions simultaneously, using a wide range of relevant knowledge and strategies" (Cumming 1990, p. 44)

We will in our study be also looking to see if differences such as these are found in the rating criteria, process, and use of knowledge sources other than the essay text itself, of raters with different amounts of training.

Interestingly, Cumming (1990) noticed that many novice teachers did, on some occasions, implement a similar range of decision-making behaviours to those that expert teachers did. In this manner, Cumming claimed that the novice teachers were on the road to developing expertise, simply as an extension of the practical knowledge of

rating which they had already acquired through evaluation experiences and teaching, but they needed time and opportunities to refine this to a point of expertise. Connor-Linton (1995) believes that this study by Cumming (1990) offers an excellent model of the rating process, especially through the analysis of think-aloud protocols, and raises serious questions about whether raters apply rating scales to essays homogeneously.

Cumming's study concludes with a comment on the diversity of features influencing the rating process, and the difficulty of controlling it (1990, p. 44):

> "The sheer quantity of interrelated decisions which occur in this process testify to the difficulty of obtaining homogenous ratings on composition exams, even among skilled raters" (1990, p. 44).

This, then, testifies to the challenge the researcher will face in this area, when attempting to discover and record all the facets of the rating process.

2.4.1.3 Vaughan (1991)

Vaughan (1991) was interested in what goes on in trained raters' minds while they are evaluating essays holistically. The call for such interest arose from the variable results in studies which investigated the reliability and validity of holistic assessment of written essays (cf. 2.2). Such studies looked at scores awarded to samples of essays purely as products, without examining the process through which these samples were rated by raters and according to which the scores for these essays were awarded.

Vaughan (1991) was another early researcher to use concurrent verbal protocols (think-aloud) as a source of data to investigate the rating process of texts produced by ESL learners. In her study, she examined 9 raters rating 6 texts, 2 produced by native speakers and the rest by non-native speakers, representing four language backgrounds.

These six essays were chosen from files from past essays at a university. The raters were experienced raters in holistic assessment. Raters received detailed written instructions and were asked to judge the essays holistically as they usually do in normal contexts: within those confines they were given the freedom to react according to what was comfortable for them. They also were guided to read through and holistically grade the six essays and verbally comment into a tape recorder as they read. The scale which the raters used was a 6-point rating scale, with passing essays receiving 4, 5, or 6 and failing essays 3, 2, or 1.

Vaughan (1991) reported that the comments made by the raters mainly focused on problems in the texts and, amongst these, content was the most frequently mentioned followed by handwriting, tense/ verb problems and punctuation/capitalization errors (consistent with Freedman (1981)). These are therefore the kind of criteria we may expect raters to rely on in our study also.

Vaughan (1991, p. 115) pointed out that the essays were given different scores from those of the original raters. The original raters passed only two of the six essays (33%). In her study, her informants awarded essays a passing grade 57% of the time. Only 44% of grades awarded were the same as those originally awarded. A further 44% were within one point. However, 12% of the remaining were within two points of the original. As already mentioned, this sort of comparison, while interesting for situations where accuracy of score is the main target of the rating, will not be within the scope of our project which targets the situation where informing the writer and assisting the learning process is the main aim of the rating.

Vaughan (1991) grouped the raters' comments into 14 general categories, six of which were mentioned in previous research, plus others that emerged in her data analysis. As

we can see, some of them refer to the usual kinds of criteria which can be applied to the essay product (A-G) and some to the process of arriving at an overall rating based on multiple criteria (I-J). H is interesting since it seems to refer to the writer's process in producing the essay product, rather than just what is in the product. It will be interesting to see if our raters refer to any such features. They are, of course, difficult to identify when a rater only has the product available to them, and did not observe the writer writing it.

Once again there are also some other kinds of comment (K-N), which do not straightforwardly map onto those identified either by Huot (1988) or Cumming (1990) above. Clearly this is the area where we may expect the greatest challenge to arise for us, since there seems to be little agreement in studies over a set of 'other comments', beyond the core evaluative ones concerned with applying criteria and deriving an overall rating/score from them. In this case, we can see for example that there are references to the rater inferring information about the writer from the text (K L). Such comments were not in the list cited above of Huot or Cumming but are intriguing. They seem to imply that the rater has identified errors which led to these inferences being made, but what role, if any, the inference plays in evaluation is unclear.

Table 2-3 Categories of comments made by raters in (Vaughan's 1991 study, p. 116)

A.	Organization	H.	Editing skills
B.	Content	I.	*"x"* mentioned as the major cause essay's failure
C.	Grammar	J.	*"x"* mentioned as the major cause of essay's passing
D.	Sentence structure	K.	Writer judged a non native
E	Coherence	L.	Writer judged a non reader
F.	Handwriting	M.	Rater responds holistically toward writer of essay.
G	Figures of speech	N.	Rater criticizes writing assessment tests.

In her study, Vaughan also focused on what she calls raters' reading style, which we would prefer to call rating style since it is clearly reading done for the purpose of rating. She provided some examples of different individual approaches to the process of assessment which are very suggestive, and will be followed up in our study, since the whole issue of individual rater's styles seems to be an important one to further explore. Vaughan (1991, p. 118,120) identifies the following:

1. The single focus approach (raters 9 and 5)

Rater 9: essay A: 'the first thing I do is I look for things that would make it a not passing essay.'

Rater 5: essay B: 'whenever it's a borderline case, I would tend to put people forward.'

2. The "first impression dominates" approach (Raters 3 and 7)

Rater 3: essay A: 'well immediately I know that this writer does not understand, is not clear, about what he is going to be writing about.'

Rater 7: 'I do not like essays that start with "I disagree with the statement …"'

3. The "two-category" strategy (Rater 4 and 6)

Raters 4 and 6 each concentrated on two categories. Rater 4 considered organization and grammar for the essays to qualify pass. On the other hand, rater 6 emphasized content and grammar.

4. The laughing rater (Rater 8)

Rater 8, seemed to establish a psychological link with the writers of the papers. This rater reacted strongly to content and got quite annoyed at one essay in particular, essay C, on attendance in college. He stated: 'well, I just do not like this student at all… essay C gets a 2 because I do not like the way it argues…'

5. The grammar-oriented rater (Rater 2)

Rater 2 reacted almost entirely to grammatical items, though he relied on coherence, in essay C: I think I'd give it a 4 because the grammar mistakes are only occasionally and it is very coherent.

Vaughan's comments about and classification of individual rating styles seen here are questionable. For instance, surely style 5 is simply an example of style 1, and laughing does not appear to be the key feature of style 4. Nevertheless, she reveals interesting possibilities, such as that some raters may, in effect, focus only on only a few criteria/traits (styles 3 and 5) (compare 2.2.3.5), or indeed use criteria which possibly were not intended to be considered in the assessment at all (e.g. how far the rater agrees with the opinions of the writer, style 4). Hence, we feel individual styles are a useful avenue to explore further in our study, and pose a research question on this.

She sees each rater as a solitary agent, making decisions mainly on the basis of observable written evidence. She stated:

> "Despite their similar training, different raters focused on different essay elements and perhaps have individual approaches to reading essays. Holistic assessment is a lonely act." (1991, p. 120)

However, arguably, the most cited of her claims concerns essays which pose problems for raters because they find difficulty in allocating scores according to the scale. She (1991, p.121) speculated that: "with these borderline cases, raters may be more apt to fall back on their own styles of judging essays."

Our study echoes the above concerns with how idiosyncratic raters may be, possibly following rather different personal beliefs about what criteria are suitable. Since in our study even less was imposed as to what scale or criteria to follow, we have a research question about this issue (choice and weighting of criteria during the rating process) which we hope will throw further light on this in a different context.

Furthermore, Vaughan concluded that if the papers are read quickly, as often happens in holistic assessment, they become, in the rater's mind, one long discourse. Apparently, this will lead to comparison between the quality of different essays, as she noted from her informants. This feature had also been noted by Cumming (1990) (see Table 2-2). Vaughan therefore draws attention to an important aspect of reading the scripts that should be taken into consideration: "the environment of the essays, the effect of papers taken as a whole on each other". (1991, p. 121). This is therefore something that we will also look for in our study.

2.4.1.4 Weigle (1994)

Weigle's (1994a, b) study, like Cumming's, also examined differences between novice and experienced raters, i.e. it investigated the effectiveness of rater training, which will be a rater variable in our study. Weigle however investigated training by actually

intervening to provide it, and hence claimed, "none of [previous studies] dealt with the issue of training" (p. 200). In her study, Weigle endeavoured to explain differences in the scores given to specific texts before and after training. Particularly, she examined how training might:

- Clarify the scoring rubric (the intended rating criteria).
- Modify rater's expectations.
- Heighten concern for inter-rater reliability. (1994b, p. 200)

Weigle's context of study was the composition subtest of the English as a Second Language Placement Examination (ESLPE) given quarterly at the University of California, Los Angeles (UCLA). The ESLPE composition subset consisted of a 50-minute essay of the student's choice, from two prompts provided (Weigle, 1994): "one prompt requires students to interpret graphical information (the GRAPH prompt) and make predictions based on this information, whereas the other prompt requires students to make and defend a choice based on information contained in a chart or table (the CHOICE prompt)" (Weigle 1994, p. 201). An analytic scale was used, with descriptions provided for each level on a range of textual features.

She observed the ratings produced by eight experienced and eight novice raters for 30 compositions responding to two task prompts. Furthermore, she recorded the process raters used while rating a selection of the compositions, both before and after the training session. Principally, her emphasis in this study was on changes witnessed in the novice raters as a result of training, and she consequently pays much less attention to the behavior of experienced raters.

While many of the findings relate to training effects on reliability and inter-rater agreement, which, as we have said, are not our concern, Weigle did identify some features of the rating process which could arise in our context. For example:

> "What can we make of the fact that more raters rate the two prompts differently after training then beforehand? One possibility is that, as raters gain experience with the two prompts, they develop separate standards for each prompt, comparing the essay at hand with many others on that prompt that they have read, rather than maintaining a set of abstract criteria in their minds that they apply to each essay, regardless of the prompt. They may refer to the guidelines less, and therefore have less of an explicit means of maintaining the same standards across the prompts. Alternatively, they may interpret the descriptors on the scoring guide differently for essays on the different prompts" (Weigle 1994a, p.108).

In our context, each teacher may be rating essays written to different prompts. Hence, we may or may not also find them working with different criteria for essays written to different prompts. We will also look to see how far they rate based on their own chosen criteria versus by comparative reference to other essays previously read for the same prompt.

Interestingly, in her study, Weigle also recognized how raters refer to sources of information other than essay scripts as a foundation for some of their judgments. For example, illustrating this, the data showed (1994b, p.213) that one rater compared her rating with what she believed a different kind of rater (a teaching assistant) would provide:

> "I know an English TA would just give this may be an F, so I do not think I should give it a 2 [in content]" (Weigle 1994, p.213)

Weigle (1994b) noticed another rater who also showed evidence of being aware of her own attitude or bias as a rater in comparison with others:

> "Okay, I think I'd probably give this one, um, a 7 or an 8 for rhetorical control. Because I'm always hard I'm gonna give it a 7."

Additionally, this rater indicated in her post interview "she felt she was a little stricter than other raters" (Weigle, 1994, p.213). While Weigle labels these comments as 'concern for inter-rater agreement' we feel that they could easily arise also where, as in our study that would be less likely to be a prime concern. One may attribute them to the rater's internal debate over what a reasonable grade could be, rather than to an attempt to agree with any particular 'community of raters' (Lumley, 2002).

Seemingly, Weigle in this study seems to confirm this interpretation when she states (1994b, p.214):

> "However, the protocol data did not show that reader agreement was an over-riding concern for raters, as Charney (1984), Huot (1990) and others have suggested as a potential danger. While all raters showed some degree of awareness of the need to agree with other raters, comments to this effect were few and far between in comparison with substantive comments about the quality of the essay itself. In fact, the few times raters did mention a tendency to be either harsher or more lenient than other raters; this was never used as a rationale for compensation for this tendency" (1994b, p. 214).

2.4.1.5 Erdosy (2000)

Erdosy (2000) worked with just four raters, chosen so as to differ in four background factors which he wanted to investigate the effects of: culture, L1, academic background, and experience as ESL teachers and learners. Each rated 60 compositions originally written by university students for the TOEFL exam. Participants were given a six point

scale, but no other information, so were free to select and apply their own criteria. He used think aloud reports to ascertain how they arrived at a score for each essay, supplemented by interviews where the participants commented on 12 of the recorded think aloud protocols.

This study therefore resembles mine in that no criteria were supplied, although I did not supply a score scale either. It differs in that a writing test scenario was replicated rather than a class writing practice one, where writers are personally known to the rater and the purpose is not pure assessment. Four raters also seemed to me to be rather few to really represent a number of background rater variables adequately, so I aimed to involve rather more raters (but fewer scripts to be rated).

Erdosy (2000) found that, depending on their background features, the raters defined writing proficiency, and the "developmental trajectory" (op. cit. p. 106) that learners would follow in acquiring it, rather differently. Hence their criteria differed. Interestingly, however, they made use not only of evidence from the essay texts, but also of their general knowledge of the writers, since they knew that they were university students. This included not just showing awareness of what sort of writing was required of the writers at university, but also of well known strategies which the writers in the context might use. One rater, for example, penalised a writer for an off topic essay on the grounds that he had perhaps learnt up model answers and was using one of these, despite it being not quite suitable to the prompt, so as to gain high marks for language accuracy. In our study, we may expect much more specific and detailed use by raters of their knowledge of the writers, since they know them as particular individuals in their classes, and not just as a general type of student.

Aside from this, Erdosy was concerned with illuminating the traditional issues of reliability and validity which, as we saw, dominate the writing testing literature. For him it is therefore seen as a failing that "rating scales are not the sole determinants of writing quality in raters' judgments" (op. cit. p. 113). If we look away from purely exam/test purposes, however, I believe that variability in rating arising from other sources, such as the raters' own criteria, their teaching experiences, and what they know about writers, might have value. They could better further the purposes which assessment of classroom practice writing needs to serve, such as informing the teaching of writing in the context, and contributing to formative feedback to the writer.

2.4.1.6 Sakyi (2000)

Sakyi (2000) studied six experienced raters rating 12 university student essays, written by students with a variety of majors, again eliciting think aloud reporting. A holistic five point scale was provided, with a very brief paragraph characterising each level, and once again exam marking was simulated.

He identified what he called four different 'reading' styles, which might be better called rating styles. One focussed on detecting and correcting language errors, and identifying the meaning of small sized units like phrases. Another was characterised by attention to essay content and organisation of ideas into text units like introduction etc. A third was voiced as the personal reaction of the rater to the language and ideas in the text. The last focused on the scoring guide which was widely cited verbatim and used to arrive at mark after reading the whole essay or, for one rater, after reading only a small part of it. This represents an insightful finding which the current study will attempt to replicate, although of course the fourth style is unlikely to be found since my study provides no scoring guide.

While Sakyi (ibid) found that raters shared a great many criteria, they tended to focus on just a few in order to arrive at the required overall mark. Especially some homed in on language criteria, others on ones associated with ideas and organisation. Their choice was influenced by their style, and their expectations, as well as what was in the essays. They also evidenced difficulty in arriving at an overall score for an essay due to the fact that they had to combine information which they had obtained on several criteria, some of which the essay maybe had done well on, some not, and no system of weighting different criteria in the final decision was provided. This again is very likely to arise in my study too.

Finally, Sakyi (ibid) refers to the notion of 'contrast'. In his study, raters often compared an essay with ones they had already read when assessing it. Thus, the order in which they read them could have influenced the scores awarded. In my study, we might expect to find this too, and indeed possibly comparison may also be made with what the same writer had written on earlier occasions, since, unlike in almost all these studies we are reviewing, in my study the teacher-raters will in fact know the student that wrote each essay.

2.4.1.7 Cumming et al. (2002)'s taxonomy of the scoring process

Perhaps the most comprehensive recent study of rating behaviour (even though it is now some 15 years old) is by Cumming, Kantor and Powers (2002) who employed think-aloud protocols to investigate the rating behaviour of ESL/EFL NNS and English as mother tongue NS raters. Cumming et al.'s study was conducted to respond to a perceived need for developing a comprehensive framework for the description of processes involved in scoring written compositions, which, as we have seen, was not evidenced in the studies reviewed above.

Based on the data they gathered, they refined Cumming's (1990) and Sakyi's (2000) frameworks for rating behaviours, establishing three general areas of the decision-making which they termed: self-monitoring focus, rhetorical and ideational focus (i.e. content and organization), and language focus. This terminology reflects the fact that, in their studies, criteria related to content, organization and language were required to be attended to by the rating task rubric, which also supplied the scale for scores to be recorded on, so focus on these areas was in a sense prompted by the rating task rather than the rater, while focus on anything else was chosen for comment by the rater autonomously, hence the term 'self-monitoring'. In our study, although this three way division may be useful, all three categories are chosen by the rater, not imposed, so in a sense all are 'self-monitoring'. What in practice distinguishes the so called self-monitoring category tends to be that this is where strategies corresponding to all those rather disparate 'other comments' that we noted in studies above are located.

Within each of the three main categories of strategy employed by raters there is a division between what is called interpretation and judgment, which appear under various names (e.g. non-evaluative and evaluative) in other classifications of rating strategies (e.g. Lumley 2005) and produce in all 37-distinct decision-making behaviours that have appeared frequently in rating think-aloud data (Cumming et al., 2002, p.72).

Table 2-4 Descriptive framework of decision-making behaviours while rating TOEFL writing tasks (Cumming et al. 2002, p.88)

Self-Monitoring Focus	Rhetorical and Ideational Focus	Language Focus
Interpretation strategies		
Read or interpret prompt or task input or both	Discern rhetorical structure	Classify errors into types
Read or reread composition	Summarize ideas or propositions	Interpret or edit ambiguous or unclear phrases
Envision personal situation of the writer	Scan whole composition or observe layout	
Judgement strategies		
Decide on macrostrategy for reading and rating; compare with other compositions; or summarize, distinguish, or tally judgements collectively	Assess reasoning, logic, or topic development	Assess quantity of total written production
Consider own personal response or biases	Assess task completion or relevance	Assess comprehensibility and fluency
Define or revise own criteria	Assess coherence and identify redundancies	Consider frequency and gravity of errors
Articulate general impression	Assess interest, originality, or creativity	Consider lexis
Articulate or revise scoring decision	Assess text organisation, style, register, discourse functions, or genre	Consider syntax or morphology
	Consider use and understanding of source material	Consider spelling or punctuation
	Rate ideas or rhetoric	Rate language overall

Table 2-4 shows that interpretation strategies involved activities which may be considered a prerequisite for rating the essays, but not actually evaluative. Raters used interpretation strategies with self-monitoring focus in reading and making sense of the input, the writing task and the supplied criteria as well as the script, and speculated on writers' personal situation. When interpreting with rhetorical and ideational focus, raters examined the rhetorical structure and summarized ideas and scanned layout. As far as language focus is concerned, raters used interpretation strategies to identify and classify errors and in their attempt to comprehend unclear portions of text.

Judgment strategies, on the other hand, are those which have an evaluative element. Some judgment strategies with self-monitoring focus related to higher level activities, above the level of evaluating particular scripts - macrostrategies for reading and rating, such as deciding sequence, comparing with other scripts, summarizing, and

distinguishing or tallying judgments. In addition, judgment strategies with self-monitoring focus concerned recording personal response, creating own criteria, expressing general impression and articulating or revising a scoring decision, in other words any judgment not based directly on the criteria (which in this study were provided). Judgment strategies with rhetorical and ideational focus concerned assessing logical structure, task completion, coherence, originality, text organization features, source use, and rating ideas. Regarding language focus, judgment strategies concerned the following: assessing quantity and comprehensibility of written texts, error features, syntax, fluency, spelling, punctuation and language overall. Overall it is clear from Table 2-4 that the number of different judgment strategies was more than the number of interpretation strategies.

A number of findings emerged other than the taxonomy itself. During their decision making stage, the experienced raters exhibited a balance between attention to rhetoric and ideas and to language features in ESL/EFL compositions that they assessed. Secondly, they also provided reasons to weight criteria for assessing ESL/EFL writing proficiency more heavily toward language aspects at the lower end of a rating scale, while balancing them more evenly between language and rhetoric and ideas at the higher end (Cumming et al., 2002, p.89). Additionally, English mother tongue raters were found to attend more to ideas than to language features of the texts. Such issues of how criteria are or are not weighted or balanced by raters relative to each other will again be investigated in our study.

This project provided empirical evidence for several issues raised in connection with rating performance. Primarily, however, Cumming et al.'s (2002) taxonomy seems to accommodate in a transparent way the essential elements of various earlier and less comprehensive lists of behaviours, and avoids the odd decisions made by Huot (1998).

Thus, we will consider using it for our own data later, even though it lacks any insight into the sequence of strategy use in the rating process (see next section).

2.4.2 Comprehensive models of the rating process as a sequence

While studies such as those reviewed above uncovered a range of strategies or comment types used by raters, both evaluative and non-evaluative, on the way to deciding on a rating for an essay, they did not directly address the issue of sequencing. In particular, as has been earlier asked in research on writing strategies and reading strategies, the question is whether raters follow a linear process or do they follow a recursive path, or both, depending on different factors (Bukta, 2007). It has also been claimed that raters go through the stages of rating a script differently, and that there is no standard order of rater behaviour: there is only one common feature: "the processes are recursive in nature" (Bukta, 2007, p.87). Thus, taking account of sequences is also part of investigating individual rating styles. Complexity of the rating process is therefore widely seen to appear in the sequences of rating behaviours, not just in what rating behaviours occur or how often.

In our study, we will pose a research question concerning this issue, so it is appropriate here to review key studies which have proposed comprehensive models of the rating process in a way that captures not just what strategies are used but also the order in which they are applied.

2.4.2.1 Milanovic, Saville and Shuhong's (1996) framework of the scoring process

One of the earliest frameworks which described the path of raters' decision making processes is Milanovic et al. (1996). The aim of this research was to find out what

occupies raters' mind from moment to moment. They compiled a model of the decision-making process built on a model developed by Cumming (1990) and based on how 16 raters of a variety of types holistically rated two different proficiency level scripts. This model therefore reflects not only the features raters attended to but also the rating process as a sequence.

Figure 2-2 below shows that the rating process they uncovered has seven stages. Each stage involves a different focus of rater behaviour. First was pre-marking, where raters internalize the marking scheme (which is assumed to be provided, unlike in our study) and interpret the writing task. This led to the scanning stage where the rater focused on surface features of the script, for instance, how long it was, text formatting, handwriting and organization. The third stage involved reading the scripts quickly and focused on overall comprehensibility. The fourth stage was the rating, which represented the focus on both the content and linguistic features of texts. "Content is assessed for text relevance, topic development, coherence and organization. Linguistic features focus on looking at errors, assessing syntax, lexis and spelling" (Milanovic et al., 1996, p.95). Next, the rater moved to the modify stage, where raters could change their evaluations of elements of the script or modify their marking focus. The next stage was to reassess or revise (including, possibly, an additional scan and/or read quickly). The final stage was the final mark decision.

Although the model is basically linear, it does allow for some recursiveness in that the rater at the reassess stage may re-enter the sequence at an earlier point. The diagram itself in Figure **2-2** does not appear to allow for anyone missing a step, however, e.g. selecting to go straight from pre-marking to read quickly. Although the model is of holistic marking, it does have built in the possibility that a rater might go through part of the sequence separately for different features, as might happen in analytic rating (e.g.

go from Rate to Final mark for content to get the content mark, then go round again for organisation). We will look to see if our data exhibits a sequence anything like this.

The model also incorporates, in the five boxes at the bottom of the diagram, lists of criteria and other aspects of the rating task that were considered at various stages, and which influence and inform what goes on in the seven steps of the core process itself. Interestingly the criterion of organisation appears at two separate points. These are essentially the components of what in 2.3.1 we called factor 2: rating task rubric, scale, criteria etc., often imposed in any rating situation but in our study mostly left to the raters to determine. It is noticeable however that the model does not explicitly mention our factors 1 and 3, that is the composition itself, and the rater characteristics, which also surely inevitably influence the way the core process progresses.

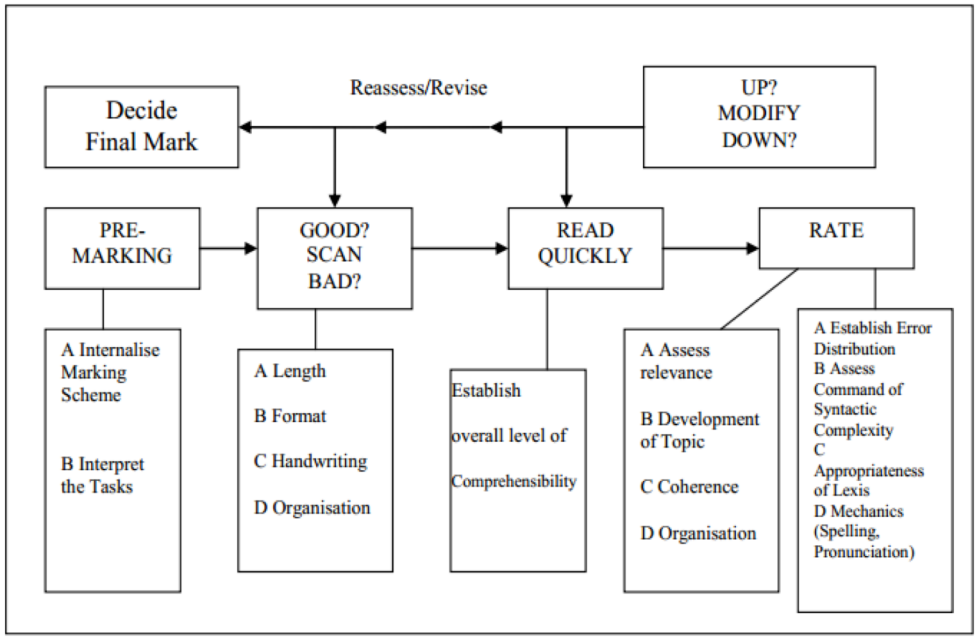

Figure 2-2 A model of the decision-making process in holistic composition marking (Milanovic et al., 1996, p.95)

The authors however also claimed in their exploratory study to have identified four reading/rating styles in terms of different ways of using the basic model, and these do in fact appear to involve missing stages or going through the whole sequence more than once. One is reading the script twice, first bearing in mind some of the scoring criteria, whereas during the second reading the focus is on determining the score, this approach they called 'principle read'. The next approach is the 'pragmatic read', which again involves two readings, but this time the second reading is done to solve any difficulties during scoring. The 'read through' approach is employed specifically with short scripts. The 'provisional approach' is the last one identified and comprises of a quick read after which a provisional mark is awarded. In this way, the proposed model seems to be so flexible that it becomes hard to see what possibilities are in fact ruled out by it.

Milanovic et al. (1996)'s findings further show while reading the scripts, raters did not reflect similar attitudes and they responded differently to the content and linguistic features. Naturally they mentioned what makes the scripts different from others or remarked on positive or negative features, and Milanovic et al. (1996) conclude that it was when assessing content that raters showed the highest degree of individual judgment.

The reasons behind these differences could be attributed to several factors of our types 1 and 3, which are not part of their model in **Figure 2-2** For instance, raters focused more on vocabulary and content with higher level scripts, whereas with intermediate level scripts, they attended to effectiveness and task completion. Furthermore, raters' background seemed to play a role in raters' differences in making decisions. In Milanovic et al. (1996) there were four raters with different backgrounds involved: raters of two levels of an EFL examination, EFL teachers and mother-tongue markers. Intermediate examination markers focused more on length, while higher level markers

focused on content and L1 writing markers focused on tone of the scripts.

Milanovic et al. claim that although they identified various differences in rater behaviour, these may not have had significant consequences. Nevertheless, these general conclusions suggest a need for further research into rater variables (1996, p. 106-107), some of which we will consider.

2.4.2.2 Wolfe's (1997, 2006) framework of the scoring process

The main issues raised in studies concerned with the sequence of steps the raters go through, according to Wolfe (1997), revolve around disagreement over whether or not different raters adopt one basic method or procedure when making scoring decisions. Wolfe (1997) was amongst others (e.g. Milanovic et al., 1996 above) who suggest that there are important variations in the rating approaches used by different raters in contrast to some who just consider that raters adopt one rating method (Freedman and Calfee, 1983; Homburg, 1983). This is then is a discussion about a different form of inter-rater agreement from that based on the scores awarded, one concerned with whether the same criteria etc. are used, and especially whether the same sequence of steps is followed in the rating process (Wolfe, 1997). In order to cope with a range of potential individual variation in sequences, Wolfe therefore designed a model with little linearity specified, in contrast with that of Milanovic et al., (1996).

Wolfe's proposed model of the rating process was arrived at not only from his own study of holistic rating, but also by summarizing the findings of a series of studies which attempted to document cognitive differences between raters who rate essays in psychometric, large-scale direct writing assessment settings (Wolfe, 2006, p. 37). He also made use of the information-processing model of holistic scoring proposed by Freedman and Calfee (1983), in which they identified three main steps that underlie the

rating of a composition: 1) *read and comprehend text* to create a 'text image', 2) *evaluate text* image and store impressions and 3) *articulate evaluation* (Freedman & Calfee, 1983, p.91). This model supposes that the rater is the one who guides the process at all stages, so of course the model may vary from one rater to another (Wolfe, 1997) (cf. 2.2.4), but rater variables are not included in the model itself, in the way that writer variables are, for example, in the Flower and Hayes (1981) cognitive model of the writing process itself.

Wolfe's model of the rating process claims to cover rater thinking/cognition in general (see figure 2-3). This model differentiates between what it calls two cognitive frameworks: a *framework of writing* and a *framework of scoring* together with the text and text image (as shown in figure 2-3). The framework of scoring is in fact the model of the rating process itself, based on Freedman and Calfee as just described, with just three broad stages of 'processing action' top to bottom, compared with the seven stages in Milanovic et al. The framework of writing is analogous to the bottom boxes in Milanovic et al., dealing with what we call factor 2, the criteria. Although it is subtitled 'content focus' it covers all types of criteria, whether focused on content or language etc. The reference to text and text image on the left captures our factor 1. As with Milanovic, there is no explicit reference in the diagram to factor 3, the rater characteristics, although these researchers clearly are aware of their impact on the processing itself.

Wolfe claims that the rater first reads the text written by the student and creates a mental image of the text (left side of the diagram). Of course, the created text images may differ from one rater to another due to environmental and experiential differences among raters (Pula & Huot, 1993; Wolfe, 2006). Next a scoring decision is made

through the performance of a series of later processing actions that constitute the framework of scoring (middle of the diagram). That is, the framework of scoring is "a mental script of a series of procedures that can be performed while creating a mental image of the text and evaluating the quality of that mental image" (Wolfe, 2006, p. 40).

For example, after reading the text in order to begin formulating the text image and commenting on the text, the rater proceeds to evaluation which constitutes monitoring specific characteristics of the text (prompted by the framework of writing), *reviewing* the features that seemed most noteworthy and then making a *decision* about the score to assign; *justification* actions which follow are diagnosing, coming up with rationale, and comparing texts. It is noticeable that in the diagram the framework of scoring box commits itself to far less detail than Milanovic et al.'s model does in terms of number of steps, and has no arrows within it, so presumably implies that the sequence may be gone through repeatedly, with omission of steps, as much as required, i.e. is fully recursive. We will be interested to see if this fits our data better than Milanovic et al.'s more detailed model.

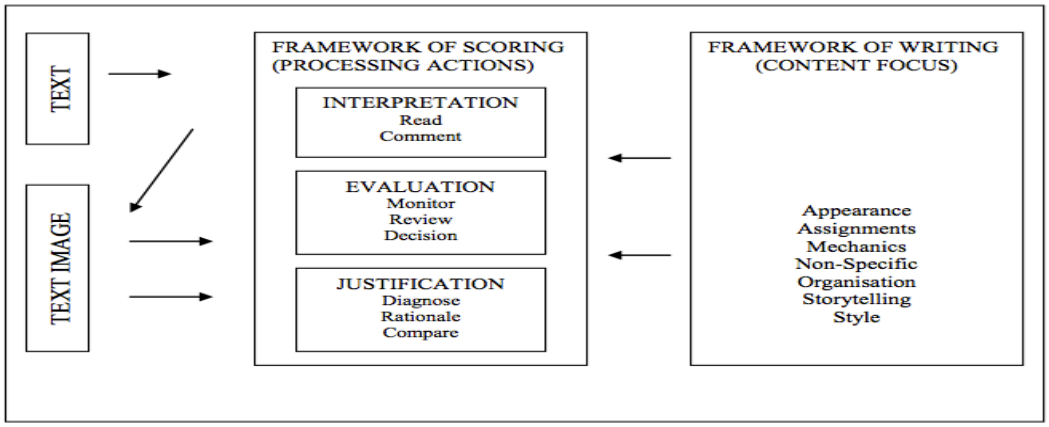

Figure 2-3 Model of scorer cognition (Wolfe, 1997, p. 89)

Besides proposing this model, Wolfe's study further tried to identify differences and similarities between raters of different rating proficiency (defined by him in terms of ability to come to agreement with other raters, rather than training etc.). He found that

differences might appear in the failure to identify the connection between ideas contained in the writing as a result of not capturing the essence of writing in the text image adequately. Furthermore, personal comments might distract the rater from the rating process.

With respect to the content focus, in Figure 2-3 above, Wolfe found that the quality of the mechanics, the organization of the student's ideas, the degree to which the student adopted storytelling devices to communicate the sequence of events, and the degree to which the student developed a unique style for presenting his or her ideas all influenced his raters' decisions (Wolfe, 2006). There is no guarantee, of course, that our raters will focus on the same features.

In addition, Wolfe found that adoption of different content focus categories (i.e., in our terms, criteria and their weighting, Factor 2) may reveal significant cognitive differences among raters. For instance, different conclusions may be reached if raters have different areas of emphasis during evaluation, such as focus on writer's style versus focus on storytelling devices (Wolfe, 2006:41). Importantly, Wolfe points out that rater differences may not be limited to the number and nature of the content focus categories used while making rating decisions, but may extend to other components of the framework of writing. For example, raters may differ in respect of the frequency with which they shift their focus and jump between content focus categories (Wolfe, 2006). In other words, individual rater styles may be characterised by the sequences of steps they follow, not just the criteria they choose to rely on.

Wolfe used think-aloud protocols to examine the jumps between categories that raters made, and concluded that less proficient raters made more jumps which suggests that they have trouble conceptualizing their decision-making process. The protocol analysis

revealed that less proficient raters tended initially to read a short section of the essay and begin to formulate a decision and then, as they read on, their decision developed. Proficient raters on the other hand tended to read the entire essay withholding judgment until the entire essay had been read. This was evidenced by the fact that the less proficient raters in this study employed more early decisions and monitoring behaviours, whereas the more proficient raters employed more review behaviours. This is something we will be interested in checking on in our study.

Proficient raters also made fewer personal comments (which by the way his model does not seem to have a place for), and Wolfe's interpreted this to mean that rating is a cognitively demanding task, and hence, if done properly, leaves no space for such comments. Conversely, less proficient raters found it difficult to cope with the task and thus they often deviated from the rating process. They also tended to focus on surface features or break the evaluation down into chunks, which ran contrary to the marking scheme provided, which was holistic marking in his study. Although we will not be using Wolfe's definition of rater proficiency, we will be interested to see if such rater differences in style are found also in our study.

2.4.2.3 Lumley's (2000, 2005) model of the scoring process

The model of the rating processes proposed by Lumley (2000; 2002; 2005) is based on the findings of his study conducted in a large-scale testing context (*STEP*)[2] with experienced raters. Four trained, experienced and reliable *STEP* raters took part in this project providing scores for two sets of 24 texts using an analytic scale. The first set was scored as in a normal examination rating session. Raters then provided think aloud

[2] *Step* is Special Test of English Proficiency used by the Australian government to assist in immigration decisions (Lumley, 2005).

protocols describing rating process as they rated the second set (Lumley, 2005, p.15). A coding scheme was developed to describe the sequence of rating strategies to establish a model of the rating process.

Findings demonstrated that raters followed a fundamentally similar rating process in three broad stages: first reading, rating/scoring and considering/conclusion. None of the scoring categories was neglected by any rater. Table 2-5 below shows the model of the core rating process which consisted of three stages broadly similar to those of Wolfe's framework of scoring and again without commitment to a detailed sequence of the type in Milanovic et al. (1996) The lack of individual style variation is no doubt due to the conditions of the data gathering differing from those of the studies above, and indeed at the opposite extreme to those which we plan to study: i.e. in Lumley's case the raters were all highly trained and experienced in what was a high stakes exam rating with highly specified criteria.

Table 2-5 Model of the stages in the rating sequence (Lumley 2002, p. 255)

Stage	Raters' focus	Observable behaviour
• First reading (pre-scoring)	Overall impression of text: global and local features	• Identify scripts • Read text • Comment on salient features
• Rate all four scoring categories in turn	Scale and text	• Articulate and justify scores • Refer to scale descriptors
• Consider score given	Scale and text	• Reread text • Confirm or revise existing scores

The first stage was the so-called pre-scoring stage, during which raters attempted to get an overall impression of the text focusing on global and local features without identifying scores. This stage included technical comments on scripts and attention was paid to surface features such as layout and handwriting (compare the first three of

Milanovic's stages). The second stage involved raters' final consideration of scale categories and focus on both text and the scale descriptors one by one to award a score. This was the actual rating, where raters allocated scores and reread sections of the text as necessary (cf. Milanovic's 4th stage). There is no mention of how the scores for the four separate categories get to be combined and assessed against each other in arriving at an overall score for a script, such as we might expect to find in our study, presumably because this is an automatic computation in the STEP exam based on the scores awarded for each category separately. The third stage was where there was sometimes consideration of the overall pattern of scores awarded, revision or confirmation, characterized by finalization of scores (Milanovic's last three stages).

The three stages which emerge here are consistent with the three-stage model proposed by Freedman and Calfee (1983), reflected in Wolfe, in which raters evaluate a 'text image' formed through reading the text itself, and filtered through their expectations, experience and background knowledge (Freedman and Calfee, 1983, p. 93; Lumley, 2002, p.255). With some exceptions, raters appeared to have similar interpretations of the scale categories and descriptors, but the relationship between scale contents and text quality remained obscure (Lumley, 2005, p.15). Raters paid equal attention to the four scoring categories/criteria provided, and each of them was examined thoroughly. Lumley (2002, p. 196) confirmed what Wolfe (1997) and DeRemer (1998) had already stated, that raters attempted to interpret the scale descriptors to match their impression of the text. Although Lumley found a close match between text and scale descriptors, he admitted that "the role of the scale wording seems to be more one of providing justifications on which the raters can hang their scoring decisions" (Lumley, 2002, p.266). This is of course less likely to be found in our study, where there is no imposed scale or criteria with any descriptors.

The model of rating stages discussed above is a simplified version of a more detailed model of the rating process suggested by Lumley (2005), which includes further features. In this rather confusingly complicated version (**Figure 2-4**) the same three stage view of the core process is retained, in this case as the three main columns of the diagram, termed: 1st reading, scoring and conclusion. However, within each stage more detailed categories are identified, with a separate three "levels at which the process operates" (p. 289) being distinguished on the vertical dimension.

The middle, so-called 'instrumental' level, corresponds most directly to what is modelled in the other models that we have considered, including Lumley's in Table 2-5. It concerns rater behaviours at the three stages, and elaborates most on the second stage, where at this level activities like balancing, comparing, interpreting and compensating are listed, and a set of sources of information or of justification is provided such as the scale categories/criteria provided, rater guidelines provided, rereading of the text, rater's own criteria etc. This elaboration means that the account moves towards including the sort of detail found in the Cumming non-sequential taxonomy (Table 2-4) although still falls far short of that.

The institutional level (top row) introduces explicit mention of what aspects of the whole rating process are provided or controlled by higher institutional agencies. While this is in principle a useful component to specify in any model of the rating process, in our study there almost no constraints of this sort, so this level will not need to be addressed. As we have said, it is the teacher not any higher agency that decides what the students write about, what criteria are used to assess it, and what scale if any to record scores on.

The lowest row of the model, covering what is called the interpretation level, is the

hardest to grasp, since it seems that the processes listed here are part of the rater thinking behaviors which are already covered at the middle 'instrumental' level. Indeed, the middle level lists a behavior called 'interpreting', so why is a separate whole level additionally needed for interpretation? And how is struggle/tension mentioned at the interpretation level different from what is called conflict at the instrumental level? Lumley's text (2005, p. 299) does not resolve these issues. An additional third level might have been better used to include in the model rater variables (our factor 3) which otherwise are still largely missing (apart from training, which is seen as institutional).

Given the above considerations, what we take from this model is only a slightly elaborated version of what we also get as a consensus from the others we have considered, i.e the idea of a sequence of three broad stages of the rating process, which can be gone through recursively and with omission of stages, maybe differently by different individual raters who possess different styles in this respect. This is what we will expect to be useful starting point when we consider sequences in our own findings.

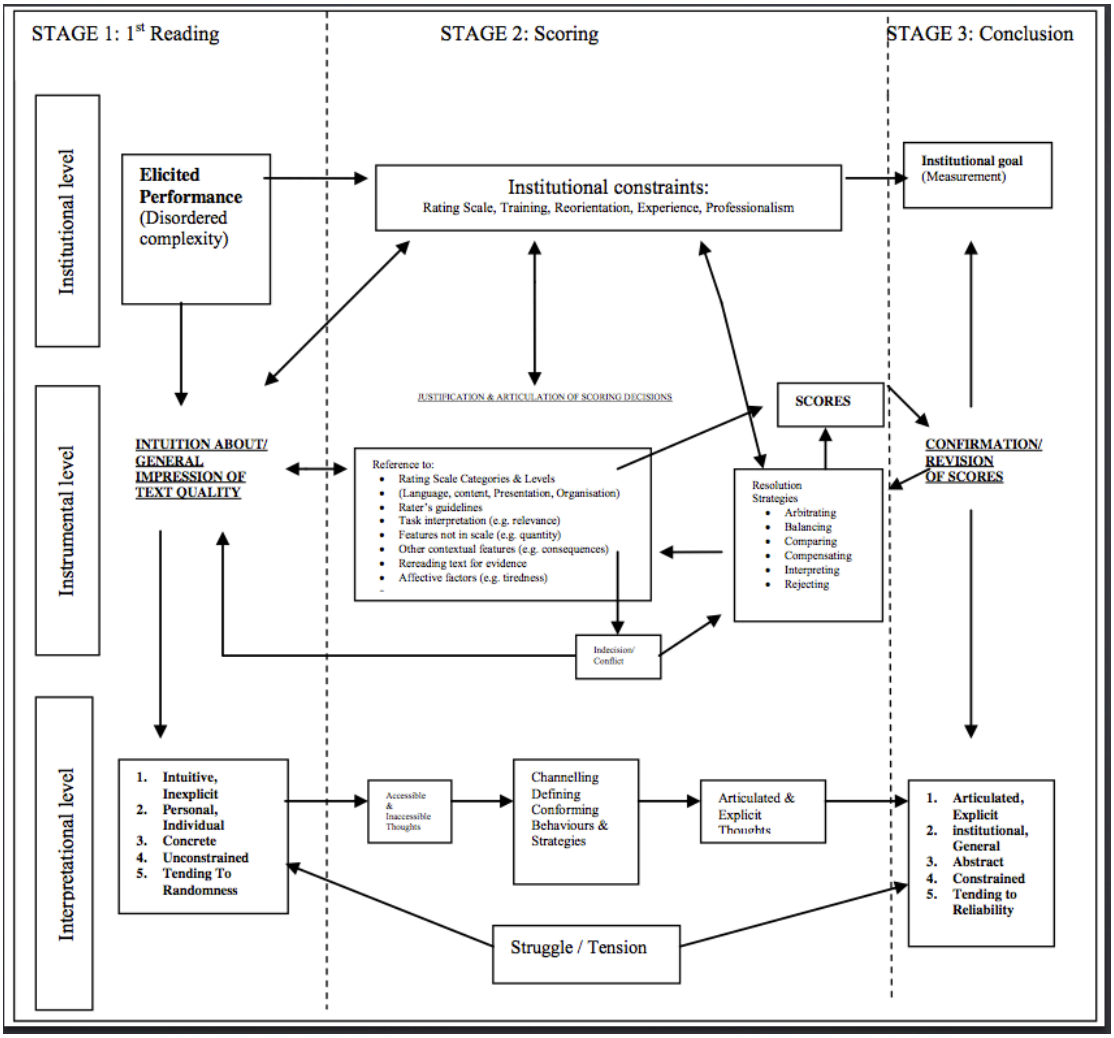

Figure 2-4 A detailed model of the rating process (Lumley, 2000, p. 289)

2.5 Instruments used to find out about the essay rating process in mother tongue and ESL/EFL contexts

A key issue in investigating the writing assessment process is the choice of a reliable and valid instrument. The literature offers a wide range of instruments, some derived from psychology (Ellis, 2008), and others from the field of SLA (Dornyei, 2005; Ellis, 2008). However, research on the essay rating process has most frequently used think-aloud reporting as the main instrument to obtain data from which to build models of the essay rating process (Barkaoui, 2010).

A search of the literature identifies 16 published studies which elicited TAPs to investigate the essay rating process. Appendix (A) lists the following information for each study: a) author(s); b) purpose of study, coded as descriptive [describe the rating process] or comparative [compares rating process across raters and/or contexts]; c) raters, type of essays, and rating scales included; d) justification provided for using TAPs; e) empirical data about TAP effects and completeness; f) cautions regarding TAP limitations; g) discussion of rater variation in TAPs (Barkaoui, 2010, p. 57).

This evidence served to confirm our intention to use think aloud as the main instrument. However, the cautions, especially concerning incompleteness, and other comments, such as use of multiple methods by some studies, convinced us that it was important to employ at least one additional instrument, so as to add to the validity of the study through triangulation. We chose immediate retrospective interviews just after the TA, and focused on it, along with questions about rating in general interviews. Such interviews were used, for example, by Lumley (2000), who conducted post-TA interviews, though they were focused on general background aspects of raters and their general views on relevant issues rather than on the rating just performed, so resemble our general interviews.

2.6 Chapter conclusion

This chapter has aimed to provide insight into research on the rating of written compositions and the factors affecting that process, which lead to variation in rating (e.g. differences in topic/prompt of composition, differences in score scale or criteria used, differences in rater training). We also reviewed relevant topics concerning the rating process, such as the kinds of behaviors observed, sequences of rating behaviors, and the TA research methods often used to gather data. Although most is written about

rating with a testing or performance assessment purpose, we attempted to focus on literature with possible implications for our study of the rating of writing done with a practice purpose, by the writers' own teachers, without any imposed scale or criteria.

In fact, in pursuing our literature review it became noticeable that only two studies (Huot, 1993; Wolfe et al., 1998) deal with rating where no scale at all is supplied to the raters, although a number supplied/imposed a scale without specifying the criteria to be used to determine how scores on the scale were to be awarded. These two studies are both now quite old, so this encouraged us to further explore this type of rating. Though not the sort of rating that would occur in a testing/examination situation, it is nevertheless a kind of rating which no doubt occurs very widely round the world when ordinary practising teachers rate student compositions written for practice. The present study attempts to contribute to this line of inquiry by examining rater behavior in the context of pre-university EFL writing courses (see next chapter).

To conclude, we feel it is useful to present our research questions with some words about the rationale of each, based on the above review.

1. What do our English writing teachers perceive as good writing? This target the criteria teachers generally believe in as important for writing, and which will come into play on any rating occasion where they are supplied with no criteria. Indeed, they may also have an influence where criteria are supplied but are unclearly stated or clash with the rater's own belief. We discussed this briefly in 2.2.4.6 where we found little literature that attempted to access such beliefs as we aim to, as a feature of the rater's belief system accessible separately from what occurs in the process of rating.

2. What rating scale and criteria do the teachers claim to typically use when rating student writing, and what are their sources? We discussed scales and criteria in 2.2.3,

but inevitably in contexts where they are supplied. We found no studies which really tackled the issue of where raters obtain criteria, other than through training or having them imposed.

3. What training have the teachers received that is relevant to assessing writing and what are their views on such training? Training emerged as a common theme in the literature, but its effect was debated (2.2.4.7). We wish to examine what the teacher raters themselves think about this.

4. What are the most important qualities that the teachers look at in practice when they are rating their students' samples? (criteria used and their weighting). This emerged in places in 2.3, in studies where these were not fully specified by the rating task specification (especially ones with holistic rating). However, we feel it is interesting to look at further and indeed compare with what we find for RQ1, for our unconstrained type of context. Furthermore, how elaborate a taxonomy will we find (cf. 2.3.1.5)?

5. How do the teachers support explain or justify their rating/scoring? This RQ covers how much and what sort of justification raters offer for the award of a mark or rating on one criterion, or when they combine individual judgments into a whole. Justification or an equivalent term occurred in almost all the models in 2.3 but does not appear to have been studied in depth. This question also entails consideration of what kinds of evidence they use (e.g. what they know about the writer, reading the text multiple times, comparing with other students...) which emerged in various places in this chapter (e.g. 2.3.1.5, 2.3.2.3). There is interest here in that some sources of evidence are available to our participants to use as justification that are not available in the usual situation of exam rating, such as knowledge of the individual writer.

6. What other kinds of comments do the teachers make, beyond those directly related to achieving the rating itself? As we saw in 2.3.1, various kinds of disparate 'other comments' have been identified in the literature, but it is unclear what a full list of such comments would contain or how far they can be seen as truly irrelevant to the evaluation itself.

7. What common sequence of activity can we find in the rating behavior of our teachers? E.g. Always read whole text first? Consider criteria in a certain order? We wish to see which of the models in 2.3.2 fits best the kind of data that we obtain in our less studied context, or whether just a three stage model is supported.

8. What distinct rating styles can we detect being used by our teachers? Rater variation per se, not necessarily related to rater experience etc., was evidenced in studies such as in 2.3.1.3 and in 2.3.2.1 with respect to choice of criteria and sequences followed. This is clearly an area that needs more illumination and is particularly relevant in a study such as ours where there is no imposed rating scheme.

9. Is there any difference in any of the above between individual teachers, according to their general training, experience of rating writing, prior knowledge of a specific rating system, etc. In 2.3 we reported studies that found or failed to find differences in rating due to such rater variables, especially training. Hence, we aim to supplement such findings with some from our study in a different type of context and see how far findings in the literature are matched

CHAPTER 3 STUDY DESIGN AND ANALYSIS

3.1 Introduction

This chapter presents an overview of the methodology, followed by the research design, participants, instruments, and procedures of the main study, the piloting, and the account of the data analysis.

3.2 Overview of the study

3.2.1 Focus of the study

As discussed earlier (chapter 2), the purpose of the present study is primarily to investigate teacher raters' reported beliefs about what is good writing, and to ascertain their rating criteria, assessment processes and decision making when they are responding to ESL classroom compositions.

Our raters responded to our questions, and also provided us with their own students' pieces of written work which they rated for writing quality. The rating was done in the absence of any scoring guide or rating scale provided by the researcher, hence these compositions were rated according to each teacher's own choice of evaluation criteria and scale. The study attempts to shed light on the process of, and factors influencing, raters' rating behavior in such response to practice writing assignments.

The rating/evaluation process does not allow direct observation, as thinking is involved, so it is difficult to design a research instrument to study it without interfering with the assessment procedure itself. We investigated raters' decision-making processes partly

by asking them interview questions and partly by eliciting their verbalized thoughts so as to reveal what occupied their minds while reading the essays: how they approached the scripts and interpreted them, how they produced the rating criteria or employed criteria they were already familiar with, and what they focused on during their rating.

The focus is not on analysing the characteristics of the text that the students produce. We are not concerned with effects on rating of actual variation either in writer ability or writing tasks and conditions. Clearly the real proficiency of the writer, and writing task-related factors such as topic or genre of writing, time allowed for writing, exam versus practice conditions, etc. may affect the writing that is done and hence the rating a student receives. These are kept constant where possible in our study. Rather our focus includes rater perceptions of these factors, which may indeed affect their rating process (2.2.1).

Similarly, we are not concerned with teacher feedback to students on their writing, except insofar as these throw light on teacher mental processes while rating. Nor are we concerned with reliability and validity of rating scales or scoring rubrics for writing, since in our study there is no one scoring rubric or system that all the raters are required to use.

The raters' thinking process is our central concern, since the score or evaluation they arrive at is primarily the result of their understanding of the performance, the assessment and the task (Hamp-Lyons, 1990). As we saw in 2.2, however, raters' criteria and processes when marking written performance may vary depending on many other rater factors, especially raters' own past writing assessment experience, previous use of a specific scale, beliefs about correctness, prior knowledge of their student writers' proficiency, individual understanding of the writing task, and so forth. These

are all examined in our study since they all contribute to making the rating process varied and complex: as raters have different experience and background they may vary in their decisions, and they may focus on one criterion more than another (Cumming et al., 2002).

3.2.2 Research approach

There is widespread recognition of at least two major approaches to conducting research in the social sciences (Mackey and Gass, 2005). One, often termed deductive or positivist, proceeds top-down: the researcher has very specific questions or even hypotheses in mind and prepares structured instruments to obtain data precisely focused on these prior issues, using categories decided upon in advance. Another, often termed inductive or constructivist (including grounded theory), proceeds bottom-up: the researcher has a general area of interest and some broad research questions, but uses open, unstructured instruments, with little control of what issues emerge from the data, and categories are found from the data during analysis. In the positivist approach, the preconceived ideas of the researcher, based on the literature or theory, are paramount; in the constructivist approach, it is the voices and perspectives of the participants that are paramount and are explored.

Having made this distinction, however, it must be said that it is often found in applied linguistic research, especially in the areas of teacher cognition and strategy research, that a judicious combination of the two paradigms is found. Indeed, this is our stance. It has already been seen (in chapter 2) that we have research questions but not hypotheses, and those questions, while not highly specific, are nevertheless not completely general either, but target some expected areas of interest such as scoring criteria, rating scale, the rating/decision making process, and effects of rater

background. Our instruments also, as will appear in more detail below, exhibit a combination of the two approaches. In particular, the data gathered is primarily qualitative, and qualitative data gathering is strongly associated with the constructivist position. On the other hand, we also use some quantitative data arising from some closed questions asked in the general interview, and secondarily from the qualitative data gathering, in the form of counts of occurrences of coded material.

In order to address the research questions, then, both qualitative and quantitative methods were used, involving basically three instruments: A. a general interview ascertaining participants' background characteristics and their beliefs about writing related issues; B. think aloud (TA) reporting to ascertain their actual rating process, criteria and scales used, and so forth employed as they evaluated essays; C. semi-structured immediate retrospective interviews conducted with raters just after each essay was rated in order to elicit more of their process of writing assessment. There was also some use made of observation by the researcher of what the raters were doing while rating, and of what they wrote on student scripts.

As Lumley & Brown (2005) have argued, interviews are commonly used to elicit views of raters and candidates about both ‘test-taking’ and rating, as an alternative or supplement to verbal reports (in our case the latter). Indeed, the use of more than one instrument to tackle the same issues, and both qualitative and quantitative data gathering (termed 'mixed methods'), is more generally viewed as strengthening a study through triangulation (Mackey and Gass, 2005; Dörnyei, 2007).

The three main instruments were implemented in a way that exhibits a balance between the constructivist approach and the positivist. The general interview contained some structured closed questions, decided on by the researcher top down, but also contained

some open response options. The instructions for the TA were largely non-directive and the interview questions only semi-structured, so this allowed the teachers' categories and views to emerge. The data analysis was partly driven by coding based on what was found in the TA data, but also at various stages informed by taxonomies of potential categories or codes from prior studies.

There is another way in which the study fits more with the constructivist approach, which is its contextualisation. While the positivist approach believes that facts can be uncovered from research which are generalisable across many contexts, and may involve research in quite artificial conditions to uncover them, with participants who are claimed to represent large populations of people, constructivists tend to believe that all knowledge is specific to particular real contexts, and specific individual people, as these are all different. The current study clearly takes the latter view. This study is designed to investigate a real course environment, where actual teacher raters evaluate their own students' compositions. The teachers responded in their own individual ways to texts from current students in their classes (papers written for assignments which the instructors had not created and assigned only for the research) while sitting in their own offices, homes, or other places in which they normally worked. After evaluation, the students got back their scripts, as normal for writing done as a class assignment rather than an exam. Furthermore, we will not claim that what we find will necessarily be true of raters of ESL writing in other contexts.

As we saw in chapter 2, most past studies in our field in fact tended to create or assume a more artificial and test-like response environment for their research, such as that of Cumming (2002) who studied the decision making behavior of experience raters using previous TOEFL samples rather than their own students' compositions. Weigle (1994), who investigated the effect of training on experienced and inexperienced raters of ESL

placement compositions in a university context, again used a previous subset of ESLPE[3] compositions written by university students.

3.2.3 Research questions in relation to instruments and data

We next present, in Table Table 3-1, the way in which the instruments and the data they yield serve to enable us to answer each RQ. This will be elucidated in detail in later sections below.

Table 3-1 shows how each RQ about the teachers/raters of pre-sessional student writing is answered.

[3] ESLPE is the subset of the English as a Second Language Placement Examination given quarterly at the University of California, Los Angeles (UCLA).

Table 3-1 RQs and the main data used to answer them

No.	RQs	Main data used to answer
1	What do our English writing teachers perceive as good writing?	General interview closed and open responses
2	What rating scale and criteria do the teachers claim to typically use when rating student writing, and what are their sources?	
3	What training have the teachers received that is relevant to assessing writing and what are their views on such training?	
4	What are the most important qualities that the teachers look at in practice when they are rating their students' samples? (criteria used and their weighting)	Coding of Judgment of Language, and Judgment of Rhetoric and Ideation from TA, and post-TA interview responses on that. Counts based on those
5	How do the teachers support explain or justify their rating/scoring?	Coding of all features of the TA, and post-TA interview responses on those
6	What other kinds of comments do the teachers make, beyond those related to achieving the rating itself?	
7	What common sequence of activity can we find in the rating behaviour of our teachers?	
8	What distinct rating styles can we detect being used by our teachers?	
9	Is there any difference in any of the above between individual teachers, according to their general training, experience of rating writing, prior knowledge of a specific rating system, etc.	Background information from general interview considered in relation to all the preceding

3.3 Teacher/rater participants

3.3.1 Rationale of selection of teachers/raters

This study, as has been discussed above, focuses on raters and the rating process, performed in a context where no defined rating scale is provided. Hence, raters are the most important aspect of our study.

It had originally been planned to conduct this study with teachers of EFL writing at Qassim University in Saudi Arabia who teach an English writing course to English majors from year three until year four. However, for reasons that will be explained with the account of the piloting (3.7), this was changed to teachers who taught writing in the UK at a relevant level (though not exactly corresponding to the original intended level in Saudi Arabia): all the teachers selected were teachers of ESL in the UK at upper intermediate and advanced levels. They taught writing either on the Essex English Language Programme (EELP), a pre-sessional program at University of Essex, or a similar writing course in the Colchester English Study Centre (CESC) institution, in the academic year 2013/2014. The reason for choosing two institutions was simply in order to obtain an adequate number of participants to conduct the study.

These institutions follow the same system in teaching ESL writing, and we selected teachers teaching at the same level, i.e. international students preparing for university study in the UK through the medium of English. The EELP course is taught in the International Academy of the university and is designed to improve general and academic English, it provides English language preparation in line with the Common European Framework of Reference for Languages (CEFR) from levels A2 to C1, and includes development of IELTs skills. CESC nearby in Colchester town provides a very

wide range of English courses, some similar to the EELP, again including IELTS preparation courses.

In any qualitative research, the main goal of sampling is to find individuals who can provide rich and varied insights into the phenomenon under investigation so as to maximize what we can learn. In our study, therefore, we had to make some principled decisions on how to select our respondents. Firstly, during the participants' selection process, we were flexible in adding more participants who could fill the gaps until we reached a sufficient number (Dörnyei, 2007). Second, we aimed to include participants who differed on key variables which our literature review showed could have an effect on rating.

We aimed to select participants, who included not only experienced raters, but also inexperienced raters, and raters who had previously been trained in some specific rating system, as well as ones who had not, so as to uncover any differences or shared processes. As discussed in our literature review, such factors have been found to have some effect on the rating of writing (Shohamy et al., 1992; Song and Caruso, 1996). Lumley (2000) also lists 'rater background' as one of three factors that influence the rating process (along with 'rating style' and 'assessment criteria'). The raters deliberately selected for this study therefore exhibited a range of background features.

Another issue was how many participants were needed. In qualitative research, Dörnyei (ibid) believes that "the sample size of 6-10 works well" (p. 127). Moreover, he maintains that well-designed qualitative research requires a reasonably small number of participants to yield rich data that is needed to understand subtle features in the phenomenon under focus. In our study, the sample size which we planned to work with was 10 teachers but for many practical reasons the final sample size was six: three

teachers withdrew from the study and one teacher proved unable to perform TAP in a relevant way. It was felt however that, taking into account the amount of qualitative data analysis that would be required, this number was sufficient to allow both similarities and differences in rater perspectives to emerge, without the project becoming unmanageable.

The type of sampling we use therefore could be described as purposive with the aim to include cases exhibiting relevant variation. However, we must admit that to an extent it was a convenience sample (Dörnyei, 2007; p. 129): beside the sampling design plan, we had to take into consideration further issues in terms of "time, money, the respondent availability" (ibid; p. 127).

3.3.2 Characteristics of raters selected

3.3.2.1 Educational background and L1

All raters held postgraduate qualifications in English language and linguistics. Most of them had graduated from British universities. The majority had British nationality and English was their mother tongue. Only one participant was from an Arab country and his L1 was Arabic (see Table 3-2). To ensure confidentiality, we refer to them with pseudonyms.

Table 3-2 Demographic information about participants

Name	T1James	T2Bob	T3Zain	T4Sam	T5Gena	T6John
Age	41-50	50+	31-40	41-50	50+	41-50
Gender	male	male	male	male	female	male
L1	English	English	Arabic	English	English	English
Years of experience in teaching EAP/ESP	18	22	6	15	25	15
Training as a teacher	Yes	Yes	Yes	Yes	Yes	Yes
Education qualifications	TEFL	BA level	PhD in Applied linguistics	MA in Linguist-ics	BA in History	PGCE
Experience/ training using a rating scale	Yes	Yes	No	No	Yes	No
Class being taught	IELTS class/ advanced level	IELTS class/inter-mediate level	IELTS class/ advanced level	Inter-mediate level	IELTS class/ advanced	Inter-mediate level
Number of scripts being assessed	5	5	7	9	7	6

3.3.2.2 Teaching experience and training

Participants of this study are a mixture of experienced and inexperienced raters. However, we cannot refer to the inexperienced raters as novice raters because they all have enough teaching experience (range six years to 25 years). They had all also had had general training in ESL teaching. On the other hand, as far as rating/assessment training is concerned, some of them can be categorized as novice raters since they had had no such training. Of the six participants, only one teacher (T1 James) was an IELTS examiner trained in the IELTS scale and criteria.

All teachers were working in a context where they were not required to use an institutionally imposed rating/scoring system. Some teachers (T2 & T5) were provided

with a rating scale from their course program but were free either to use it or not. Two other teachers (T2, T3) also sometimes use IELTS criteria but are not examiners and have never trained to use IELTS criteria. Two others (T4 & T6) claimed to have their own designed scale that suited the genre and the students' level of proficiency. More will be revealed in the Results (5.3, 5.4) about the participants' choice of criteria and scales, and their training.

3.3.2.3 Gender

We included in this study a mixture of both male and female participants, although gender of raters was not considered to be an important factor. O'Loughlin (2000, 2002) for example, in a study of the IELTS oral interview, revealed that the gender of raters (as well as candidates) did not have a significant impact on the rating process. He concluded that 'gendered differences are not inevitable in the testing context', and 'gender competes with other aspects of an individual's social identity in a fluid and dynamic fashion [in particular contexts]' (O'Loughlin, 2002, p. 190).

3.4 Composition scripts to be rated

3.4.1 Rationale of selection of scripts

In this study, it was felt that it would be more realistic, and so more valid, if the informants rated their own students' scripts, as they would normally do for a class practice writing assignment. Using such scripts was expected to maximize the possibility of finding factors based in rater perceptions of the student, as well as of the text produced, which might influence the raters' rating, for instance the rater's idea of the students' level, knowing the student's name and previous performance, etc. Such

student factors are of course usually filtered out in studies of rating exam writing since raters in those conditions usually have no way of knowing the writer's identity.

3.4.2 Selection of scripts and their characteristics

Composition scripts were chosen written by ESL students who studied English language for academic purposes from the two English Language teaching centres. Both institutions offer English language courses for different proficiency levels. The scripts were selected by the teachers from their own classes following the researcher's instructions, which were to provide the researcher with at least five scripts each. Some teachers however were able to provide us with more than five some did not. The scripts represent different ranges of writing proficiency as it was difficult to find an adequate number of teachers who teach the same level in two institutions.

The scripts varied in their tasks as they represent two proficiency levels (CEFR B2 and C1). The level of difficulty was varied according to the writers' proficiency level. Two teachers had scripts on one of two topics each, while four teachers each had scripts with only one topic. Some topics were descriptive, some argumentative so vary in genre. Finding the same genre for all teachers was a challenge and could not be fulfilled for this specific study. Examples of topics/tasks are:

- In many countries children are engaged in some kind of paid work, some people regard this as completely wrong, while others consider it as valuable in working experience. Do you think that children can work and get paid?
- The internet enables students today to be better informed than the students of fifty years ago. To what extent do you agree or disagree with this statement? Write 250 words.

- Obesity has become an epidemic in many countries. What are its causes and can it be solved?
- Younger and older people using technology: do they use it in different ways?

Some of the scripts were typed the others were handwritten. The proposed length for the scripts varied between 80, 150 words to 250 words. Some of the topics had prompts or a rubric for the writer to follow in order to fulfil the task requirements.

There were no conditions imposed by the researcher on how to organize or to gather the scripts, i.e. this was left in the teachers' control. It was not feasible in any case to set conditions for the teachers regarding this issue. The time of the study was limited so it was not applicable to wait in order to gather a highly matched set of scripts from each teacher, written by learners from the same proficiency level and representing the same genre and written in uniform conditions. The main stipulation was just that each script was to be from a different student in their class, written for a recent class practice writing assignment (not a test/exam).

The writers of the scripts were planning to study in various disciplines but they were taking the same classes. They represented a wide range of L1s with different cultural backgrounds and a wide range of nationalities. Each teacher however had a similar mixture of learners.

None of the scripts were to be processed by the teacher before the study, i.e. given any previous feedback or rating. One teacher in fact did not adhere to this and in the think aloud sessions showed evidence of having already read over the scripts before meeting with the researcher. Nevertheless, it was considered too late to exclude him and try to find a replacement teacher at that stage. The total of the scripts was 39 scripts (see

Table 3.2). For the TA sessions, we used all the scripts that each teacher provided without excluding any.

3.5 The instrumentation

3.5.1 Introduction to the instrumentation

In order to meet the primary objectives discussed earlier, data for the study were obtained from three main introspective instruments: a general interview with a mix of closed and open response items, concurrent verbal reports (think aloud or TA) and immediate retrospective interviews after the TA rating of each essay. These were supplemented in a minor way by the researcher observing the teachers while they performed the TA tasks and making notes, and looking at the scripts they rated and anything they wrote on them.

3.5.2 General interview

This instrument included initial questionnaire-like questions where participants were asked to report on a set of demographic and other personal background features which we needed to know about for the study (as reported under Participants above). There were also both closed and open questions designed to find out mainly about three areas: how important they thought various qualities were in contributing to a good piece of writing; what scales and criteria they typically used and where they came from; any training received relevant to rating and attitudes to such training. These areas were asked about in general terms, not specifically about the scripts they were about to rate, so the information obtained arguably reflects their general beliefs and may differ from that arising from the TA while they were performing an actual rating task.

General retrospective instruments such as this are widely used to gather background information about participants (Dörnyei, 2007). With respect to other kinds of information, they are widely used to obtain reports not only of facts but also, as in our case, beliefs and attitudes which are not open to direct observation. Our general interview was quite structured and included some closed items reminiscent of questionnaires, where participants responded to each of a set of items on a rating scale offered (e.g. concerning possible features of 'good writing'). This yielded quantitative data. In the open response items participants were asked fewer more general questions with no choices and simply said whatever they want in answer to the questions, yielding qualitative data. Commonly claimed benefits of structured interviews / questionnaires are that one obtains parallel information from all participants since they all answer exactly the same questions, that they can be used to quickly obtain information from a lot of people, and that the data obtained is quicker to handle. In our case, although we were not concerned with a lot of participants, we needed to obtain parallel information quickly and easily in this phase of the study since the other two main instruments were quite time consuming and demanding for the participants.

Disadvantages of structured interviews include that participants may not be truthful in their replies and that although the researcher may be present to resolve any queries which the participants may have about responding, there is little real interaction between participants and researcher that allows for exploration beyond the questions asked. In our case, we hoped to overcome these by administering the questions orally and with each participant separately face to face.

Similar instruments have been used in our field for example by Cumming et al., (2002) and also Khongput (2014) who we based ours on.

3.5.3 Think aloud reporting

3.5.3.1 Nature of think aloud reporting and its advantages

Information on raters' mental processes when selecting and applying their criteria to reach their final rating was obtained from concurrent think-aloud verbal reporting. The think-aloud protocols (TAPs) which emerge from this method yielded qualitative information concerning cognitive processes, which is regarded (following Ericsson & Simon 1993, p.30) as more immediate than that from retrospective interviews and unmediated by unwanted factors such as memory failure or a deliberate agenda of the participant to project a false impression. From such information, as Erdosy (2000) points out, one could supplement the direct evidence of the scores or other ratings which participants assigned to compositions with evidence produced by participants' introspection, which was further tapped in the course of the immediately following interviews.

Think-aloud reporting performed in real time while doing a task requires a participant to verbalize 'what is going on through their minds as they are solving a problem or completing a task' (Mackey & Gass, 2005, p. 79). Dörnyei (2007), based on the discussion by Ericsson (2002), reports that this method involves the 'concurrent vocalization of one's "inner speech" without offering any analysis or explanation'. He also points out that the method 'is not a natural process' and thus 'participants need precise instructions and some training before they can be expected to produce useful data'. Researchers employing this method therefore need to provide participants with preparation for the tasks (p. 148), as we did (see 3.6.3).

The principle advantage of concurrent think-aloud reporting in the context of my study is that it provides evidence of cognitive processes that is not coloured by deep

introspection (Ericsson& Simon 1993, p.30). As Green (1998, p. 4) points out, concurrent think-aloud protocols do not report straightforwardly on deep cognitive processes, but they do capture more superficial thoughts from which such processes may be deduced.

In the present study, such evidence is really useful in allowing us to detect whether raters paid attention to a similar range of textual qualities in assessing compositions. In other words, it guided us to see whether the raters have the same construct in mind when assessing 'writing quality'. Erdosy (2002) considers that such evidence can be analysed both qualitatively and, once coded, quantitatively. Indeed, that is widely true of qualitative data of all sorts including that from interviews as well, and we will do this.

Another advantage of using think-aloud reporting is the sheer quantity of data, as seen in Cumming's study (1990) where the raters of ESL compositions averaged 32 coded comments for every composition they rated (See chapter 2). Hence, it is not surprising that think-aloud reporting has been widely used in investigating writing assessment.

3.5.3.2 Criticisms of think aloud reporting

This is not to say that this method is without drawbacks. Like any other instrument, think aloud protocols have their limitations. Firstly, think aloud reporting is difficult to administer because participants are often not used to verbalizing their thoughts while focusing on the completion of a task (Smagorinsky, 1994). Another difficulty is that the informants verbalize their thought processes to differing degrees (Erdosy, 2000). Additionally, the method has been found to be time consuming and labor-intensive in gathering, transcribing, coding, and analyzing data (Green 1998; Smagorinsky, 1994).

Hayes, & Flower (1980) pointed out three major objections to the use of verbal reports as data:

- They are unreliable because people are not conscious of their cognitive processes.
- Reporting the processes verbally distorts them.
- They are incomplete and not objective.

Hayes & Flower (1980) however made extensive use of TA themselves, arguing that it was strange that TAPs "are singled out for criticism on the grounds of incompleteness, since they are characteristically more complete than most other methods to which they are compared." (p.xx). Indeed, we could extend this argument to point out that other methods of gathering qualitative data, such as interviews, open questionnaires and diaries, are also no better than think aloud with respect to being subjective, open to distortion due to how thought processes are reported, and unable to capture cognitive processes that occur at the subconscious level. These others also suffer from the relative disadvantage of usually gathering their data at a time more distant from the moment when the participants actually perform the processes being reported on than the time when concurrent TA does. Hence, they are more prone to forgetting.

In this research, we pay due attention to the limitations of verbal protocols, so we carefully:

a) Prepared clear instructions (written and/or recorded) and trained the informants to generate think aloud protocols prior the main task.
b) Admitted in our interpretation of the research results that such data represent only information that people attend to and are able to verbalize during the task performance (i.e., not their full cognitive process which cannot be verbalized and

c) Avoided or acknowledged problems of reactivity or other unintentional influence on participants' behavior (Smagorinsky 1994; Cumming et al., 2002).

3.5.3.3 Use of think aloud in previous rating studies

Despite some opponents' criticisms of using the method, many proponents of concurrent verbal reporting or think-aloud (Olshavsky, 1976-1977; Ericsson and Simon, 1984/1993; Afflerbach and Johnston, 1984; Olson, Duffy & Mack, 1984; Cohen, 1987; 1996; 1989; Afflerbach, 1990; Pritchard, 1990; Wade, 1990; Wade, Trathen & Schraw, 1990; Baumann, Jones & Seifert-Kessell, 1993; Matsumoto, 1994; Pressley and Afflerbach, 1995; Klingner, 2004; Pressley & Hilden, 2004; Koda, 2005; White, Schramm & Chamot, 2007) have illuminated the advantages gained in utilizing the instrument to unravel the more or less conscious processes that lie hidden in raters' cognition and stressed its greater usefulness than that of other research instruments.

In particular, there are a number of substantial studies of rating which have employed introspective techniques, particularly concurrent think-aloud reporting in language assessment, in the past 20 years. Vaughan (1991) was one of the first using concurrent think aloud protocols in investigating the rating of texts produced by ESL learners. In her study, Vaughan (1991), describes how she 'borrowed a technique used by Raimes (1985) and others to elicit writers' thoughts: the think aloud protocols.' (1991, p.113). In her study, she justified the validity of using think-aloud reporting of the rating process by reference to 'Raimes cites Ericson & Simon's (1980) review of the literature tracing the mental processes,' (Vaughan, 1991, p.113).

Moreover, Cumming et al. (2001, 2002), Lumley (2000), and Smith (1998, 2000) used the method in investigating scoring decisions while rating writing tasks. Sakyi (2000) believes that TA verbal protocols provide direct evidence about thought processes

which complements the data of correlational studies where researchers usually analyse essay products for traits associated with high or low scores.

This approach has proved that it is capable of producing rich information about rater behavior and other factors related to the holistic scoring of written composition (Cumming1990; Hamp-Lyons 1991; Vaughan 1991; Hout 1993; Janopoulus 1993; Weigle 1994; Smith 1998, 2000; Zhang 1998; Erdosy 2000; Sakyi 2000; Cumming et al 2001). Sakyi (ibid, p.130) maintained that the "verbal protocol is used to obtain insights into what actually goes through the raters' mind when they rate compositions, particularly key decisions and attention to specific criteria for making judgments".

Lumley's study (2005), more similar to ours, is also indicative, since based on previous studies which he examined in his research, he pointed to the value of TA to shed light on the following aspects of the rating process, and the related questions that arise, many of which we are also targeting:

- The assessment criteria and how raters use them.
- The role and adequacy of the rating scale if any is used.
- The importance of rater background and experience.
- The role of rater training
- The scoring strategies raters use.
- The rating style adopted by raters.
- The role of other influences.

We expect many of these to be clearly evident from the data of our think aloud protocols.

3.5.4 Immediate retrospective interviews

3.5.4.1 Nature of retrospective interviews and their advantages

Generally, "an interview is a conversation, usually between two people... where one person – the interviewer – is seeking responses for a particular purpose from the other person: the interviewee" (Gillham, 2000, p. 1). In qualitative research, Rapley (2004) adds that interviews are 'social encounters where speakers collaborate in producing retrospective (and prospective) *accounts* or *versions* of their past (or future) actions, experiences, feelings and thoughts' (p. 16). This characterisation suits our retrospective interviews more than our general interviews, since the former were quite open and contingent upon what had been heard in the immediately preceding TA, while the latter were quite structured. Moreover, an in-depth interview, which usually consists of open, direct, verbal questions, is the method used when, as in our study, 'the focus of inquiry is narrow, ... the respondents are familiar and comfortable with the interview as a means of communication, and the goal is to generate themes and narratives' (Miller and Crabtree, 2004, p. 189). In the same vein, Hesse-Biber and Leavy (2006) note that this method is useful when 'the researcher has a particular topic he or she wants to focus on and gain information about from individuals' (p. 120). They also emphasize that in-depth interviews are 'a meaning-making' and 'knowledge-producing conversation' that occurs between two parties. In qualitative research, interviews are often used to provide such in-depth insights and better understanding of the phenomenon or behaviours(s) under observation in the participants' normal contexts (McMillian & Schumacher, 1989, p. 119).

In applied linguistics research, in this type of interview, while the researcher tries to ask each interviewee a certain set of prepared questions, he or she also 'allows the conversation to flow more naturally, making room for the conversation to go in new

and unexpected directions' (Hesse-Biber and Leavy, 2006; p. 125). This is also often termed a semi-structured interview. Indeed, Dörnyei (2007, p. 136) states that most interviews conducted are the semi-structured interview and it is suitable when:

> The researcher has a good enough overview of the phenomenon or domain in question and is able to develop broad questions about the topic in advance but does not want to use ready-made response categories that would limit the depth and breadth of the respondent's story.

In an immediately retrospective interview, according to Dörnyei (2007), "the respondents verbalize their thoughts after they have performed a task or mental operation" (p.148). In this regard, Dörnyei (2007) argued that, to retrieve more valid retrospective protocols from memory, they should be retrieved shortly after the verbal reports. Gass and Mackey (2000) further pointed out the importance for the researcher to transcribe and analyze the respondents' TAP to make the retrospective session really meaningful and the time lapse should not exceed two days and should preferably be less than 24 hours.

3.5.4.2 Rationale for use of interviews in the study

This study employed immediate semi-structured retrospective interviews with the subjects, which constituted the second main instrument in this study. As pointed out in reading research (McDonough & McDonough, 1997), relying on verbal reports drawn from learners' reading behaviours without post-task interviews with the readers might yield insufficient data as to why and how some strategies have been employed in processing the text(s). Hence, retrospective interviews help compensate for the low-awareness introspection of concurrent think-aloud reporting, used to minimize the disruption of the task, and can thus add a high-awareness dimension to participants'

reports. Urquhart & Weir (1998, p. 247) argue that ascertaining readers' purposes for reading texts can, for instance, be established by having them interviewed on how they have performed the task. Similarly, this study used retrospective interviews after the participants produced think aloud protocols, so as to find out in greater depth about the scoring/rating themes which we are interested in.

3.6 Data collection materials and procedures

3.6.1 Overview of the procedure

The main steps of the data collection were: ethical preliminaries, the general interview, TA training, and the TA rating tasks of multiple student essays, each followed by a retrospective interview. All had been piloted before the main study (see 3.7). Each step involved both preparing some materials and a procedure of administration.

After the ethical preliminaries, data for this study were then collected from December 2013. Each rater individually attended data collection sessions. The IA teachers attended one of the English Language and Linguistics Department offices, while we gathered data from the CESC teachers in an office in the CESC institute. Originally the data was planned to be gathered in a single session of not more than two hours. However, for practical reasons, this usually became two sessions. Most of the teachers preferred this because it was difficult for them to think aloud for a long time. Furthermore, the teachers were busy and it was difficult for them to set aside two to three hours at once. Most of the teachers attended two sessions on two different days and in each session evaluated a number of scripts and provided immediate retrospective responses for each script that had been rated. Just one teacher managed to produce all the TAPs and the retrospective interviews and answered the general interview in one session.

3.6.2 Research ethics

This research project rigorously followed UoE ethical guidelines throughout its process. Consent forms including a brief description of the nature of the study were distributed among the teachers before any data was gathered (see appendix B). The teachers read the forms and signed them if they agreed to participate. At the same time, I reassured them about confidentiality and that they have the right to withdraw from the study at any time. There was one teacher who withdrew before the beginning of the training task and two teachers withdrew after doing four think-aloud protocols. Hence, they were excluded and their data were discarded. To secure confidentiality, in our write-up, raters will be referred to with ID codes consisting of T+ a number 1 through 6 and a pseudonym.

3.6.3 The general interview

This instrument was delivered once, usually before the beginning of the think aloud sessions, for each teacher. It comprised a number of sections covering: background information/demographics, teaching experience, training experience on assessing writing and attitudes to training, views on good writing, writing in a composition/essay course, criteria and scales used in marking compositions/essays and other issues about writing assessment (see the full questions in appendix C).

The first three of those sections gathered important background information we needed to know so as to be able to talk later about how individual differences between teachers might have affected their rating. The other sections were designed to collect additional kinds of data on the reported writing assessment practices of teachers, and related issues, of a different type from that obtained through the TA (3.5.2).

Regarding their views about good writing features, the teachers were presented with flash cards with a list of seven common writing features and were asked to state the degree of importance for them of each of the seven features by rating each on a five point scale from (not important) to (very important). The features asked about were: cohesion (unity of ideas and structure), task completion, relevant development of ideas, appropriate grammar, appropriate vocabulary use, organization, mechanics (e.g. spelling, punctuation, capitalisation). These were chosen based on common key features in the literature referred to as contributing to effective writing, both at discourse level (such a communicative effectiveness, task fulfilment, register and organisation of writing), and at a linguistic level (for instance linguistic range, lexis and accuracy) (Cumming et al., 2002; Hawkey & Barker, 2004; Barkaoui, 2007; Knoch, 2011).

3.6.4 The TA training / warm-up

The think aloud technique was described to the raters and they had an opportunity to practice it, since it is widely recommended to provide the participants with training to familiarise them with the TA instructions and how to handle the process of a TA task (e.g. Bowles, 2010; Brown & Rodgers, 2002; Dörnyei, 2007; Nunan, 2012; White et al., 2007).

This study considered several training tasks in order to choose the appropriate one. There are two types of warm-up tasks commonly used: arithmetic and verbal. An arithmetic task presents the participants with mathematical problems such as adding or multiplying and asks them to think aloud while they perform the task. Green's model task (1998, 47-48) is of this kind: participants were asked to first think aloud and then retrospectively report what they were thinking as they were adding up all the windows in their houses (See Appendix E for the task). A verbal task on the other hand, according

to Green (1998), is exemplified by solving an anagram which is "a word or phrase whose constituent parts have been rearranged, resulting in 'nonsense' words" (Nunan, 2012, p. 118).

Both tasks were used in the pilot study but proved unsuitable (see section 3.7) so the present study finally opted for warm-up/training with a real composition from the set selected by each teacher. The researcher did not comment on their performance, so as to avoid prompting them in what to say, and data from this was not analysed. This technique was adapted from Lumley (2005), in which the participants were provided with a sample of students' written scripts. The purpose of this was twofold: it filled the place of training recommended by Simon & Ericsson (1984, 1993) and hopefully would ensure that the participants could carry the required task. Using such a technique "seemed logical ....., rather than something unrelated such as mathematical problems proposed by Ericsson & Simon" (Lumley, 2005, p123).

3.6.5 The TA verbal reporting

3.6.5.1 The verbal report instructions

The participants were provided with clear instructions for the think aloud task following the sample instructions provided by Green (1998, p. 46-47), based on Ericsson & Simon (1993), and by Lumley (2005). These instructions also followed the guidelines provided by Bowles (2010) for better instructions:

> Minimally, this set of instructions should include (1) a description of what is meant by "thinking-aloud," (2) the language(s) participants are allowed to use to verbalize their thoughts, and (3) the level of detail and reflection required in the think-aloud. (p. 115).

The sample instructions given by Green were in line with these criteria although the 'language' criterion (2) was not specified in Green. Sanz et al. (2009) did not mention

the language(s) that participants were allowed to use in the think aloud because their learners were monolingual native English speakers. Likewise, this feature was largely redundant for our instructions as all informants used English either as L1 or L2. Teachers, particularly those accustomed to communicative language classrooms, may in any case assume that they are to speak their thoughts in the language they normally use if they are not otherwise instructed (Bowles, 2010; p.115). Interestingly, one informant whose English is L2 used it in his think-aloud aloud instead of his Arabic L1 without being instructed.

For the sake of obtaining rich information and discovering details of the rating process it was important to make a decision on the level of detail (3) requested by the instructions. Lumley (2005, p. 119) for example refers to questions requiring deep awareness by the rater of what they are doing, such as 'explain why you give the score you give', 'why do ignore this aspect', 'what does this mean'. However, we preferred to adopt the forms of verbalization which were characterized by Ericsson & Simon (1984, 1993). In the type which we adopted, verbalization takes the form of subjects vocalizing all their thoughts that pass through their heads naturally while they perform the task without analyzing them. In this way, the emphasis is on the performance of the task with the verbalization as an incidental activity. The subjects in this type of TA are trained not to add other information in the form of explanation or justification (Ericsson & Simon 1984, 1993; Green 1998). Our aim was to reassure to the participants that we wanted their thinking, not what they think about what they are thinking (Brown & Rodgers; 2002).

Our actual instructions were as follows, provided both verbally and in written form. (See Appendix D)

Please read these instructions through carefully before you begin the assessments.

Purpose

These instructions are written to help guide you and others producing think-aloud protocols for this research, in a consistent and informative manner. Think-aloud protocols ask people to say everything they think about while they perform a task, with the aim of documenting and better understanding what you pay attention to and consider important when you do a task. The purpose of the think-aloud protocols for this study is to find out in as much detail as possible what you, as a teacher/rater of ESL compositions, are thinking about, deciding, and doing while you rate a sample of ESL compositions. The most important thing to emphasize is, say everything you are thinking about, and make certain this is recorded clearly onto the recorder. What you say will become important data for our research.

I would like you to talk and think aloud as you rate your students' written compositions while this recorder records what you say.

First, you should identify each script by the ID number at the top of the page and the task.

Then, as you rate each script, you should vocalize your thoughts.

It is important that you keep talking all the time, registering your thoughts all the time. If you spend time reading the scripts, then you should do that aloud also, so that I can understand what you are doing at that time. In order to make sure there are no lengthy silent pauses in your rating, I propose to sit here, and prompt you to keep talking if necessary. I will sit here while you rate and talk. I will say nothing more than give you periodic feedback such as 'mhm', although I will prompt you to keep talking if you fall silent for more than 10 seconds. Thanks!

3.6.5.2 The rating task

The actual task was rating seven of their own students' written compositions, in whatever order they chose.

Following Cumming et al, (2002), Cumming (1990), Connor-Linton (1995b), Kobayashi and Rinnert (1996) and Erdosy (2004), in this study, no criteria nor analytical categories nor rating scale nor rules for arriving at an overall rating from ratings of separate features were provided, with the purpose of finding out what scale and criteria these teachers followed themselves. In fact, this was quite realistic since, when the researcher asked the course director about the scale that is used for the EELP course, he stated "no rating scale is provided; this is among the issues that I will investigate in the near future and might include one" (personal communication 2013). No scale seemed to be imposed in CESC either.

Following Brown & Rodgers' (2002) suggestion about avoiding being too directive in the instructions to the participants, the researcher did however give some general instructions about the rating:

- Assess these scripts in the way you normally do.
- You are free to select what script to start with.
- You are free to give feedback as in your normal practice essay situation.
- You are free to write or draw on the script if this is part of your feedback practices.
- You are free to provide a score or not.

After this, all the informants followed their normal response behavior without being controlled by the researcher's instructions or prompting shaping their responses.

3.6.5.3 Presence of the researcher during TA data collection

The researcher's presence in the room is considered to be an important issue in data collection. The researcher was in the room while the raters produced the TA protocols, since as Ericsson & Simon (1993) have suggested, it is necessary for the researcher to be there to prompt the subjects if it is necessary, if there is a pause in the talking exceeding 10 to 15 seconds. Likewise, Green (1998) emphasized the need to inspire the subjects by asking them to 'keep talking' or 'can you tell me what you are thinking?' if they fall silent for a period of time (p. 42). The Pilot had shown that such silence can happen.

In this sort of TA, the researcher is invisible to the participant (behind the subject or at the corner side of an L-shaped table) and her role is to provide no more explicit prompt than 'keep talking' or 'think aloud', when the subject falls silent during the performance of the task. In the present study, the researcher sat on the long arm of an L-shaped desk whereas the teacher sat on the short side.

The TAPs and following retrospective interviews were recorded using two digital recorders. One was a Sony digital recorder with new battery which was tested twice: before the beginning of the task and at the beginning of the task. The other device was an iPhone 4 which accepts voice memos. It is an important arrangement to have a spare recorder in case the main recorder went wrong or the sound quality was weak. Meanwhile, the researcher took some observation notes. These notes made were for two reasons: the first reason was to make the participants feel relaxed and not observed or looked at, as would occur with video-recording. Secondly, these notes might yield some details that might not be captured by recordings.

After we finished recordings we made copies of the original recordings by backing up the files.

3.6.6 The retrospective interviews

These interviews occurred repeatedly for each teacher, following immediately after each TA for each script, and focused on the script just rated, which was still in front of them. It was recorded in the same way as the TAP. It is important to clarify that the retrospection referred to the scripts, not to the recordings of the TA which were not replayed.

Although, Gass and Mackey (2000) pointed out the importance for the researcher to transcribe and analyze the respondents' TAP to make the retrospective session really meaningful, in our case it was not feasible to follow this as the teachers were busy and could not fit in another time for retrospective interviews. Hence, we opted to conduct the retrospective interviews immediately after the think-aloud task for each script, before analyzing it. Thus, we aimed to combat potential problems related to issues of memory and retrieval by following Mackey and Gass' (2005) recommendation that 'data should be collected as soon as possible after the event that is the focus of the recall, the stimulus should be as strong as possible to activate memory structures, the participants should be minimally trained, and the level of structure involved in the recall procedure is strongly related to the research question' (p. 78 - 79).

Retrospective Interviews were suitable for the present study largely because they supplemented the already gathered TA evidence. The retrospection further allowed the participants to explain, rather than merely report on, their behavior and thought processes, something which they had been explicitly requested to avoid in the course of producing their think-aloud protocols. In our study the retrospective interview

therefore added to validity by being used to triangulate data obtained from the think aloud reporting.

The retrospective interview questions in the present study was designed to elicit information in a manner which was controlled but also flexible in responding to the substance of each teacher's think-aloud data. Thus, the interviews were semi-structured in that there were a few set questions, written in advance, with many follow-up questions based on teacher responses to those. They were designed to focus primarily on aspects of rating relevant to our research questions. This technique helped the researcher to ascertain more about the rating criteria used in practice by the teachers, and the process which the participants followed in order to reach the final rating decision (especially relevant to answering RQs 4-8).

The list of questions which we drew on in the retrospective interviews included the following, not necessarily in the precise wording given here (see example in appendix K):

- What grade would you give this if you were required to?
- What makes this essay good?
- What makes it bad?
- What are the criteria that you looked for the most? Do you weight the criteria the same?
- What are the factors that affected your decision?
- Tell me about what you wrote on the script and why?
- Tell me about your reading and rereading practices?
- How much do you relay on first impression?
- Do you use different criteria for different students?

3.7 Piloting the study

3.7.1 Conduct of the Pilot

It is always recommended to pilot any study (e.g. Bowles, 2010). The literature defines the pilot study as "a small scale trial of the proposed procedures, materials and methods and sometimes also includes coding sheets and analytic choices" (Mackey & Gass, 2005, p. 43). The pilot study was an insightful stage where I tried out the three main instruments I planned to use and obtained feedback on them, since the aim of a pilot study is find out any flaws in a planned study and address them in advance (Mackey & Gass, 2005), which helps save time, effort and money.

The Pilot was conducted in January 2012 with five teachers of EFL composition writing to first year English majors in the English Language Department at Qassim University, which was the context where we initially planned to carry out the main study. The five teachers who agreed to take part in the research were in fact all Pakistani, with post-graduate qualifications, and extensive experience. They each responded to the general interview and provided think-aloud protocols while rating three scripts written by Saudi EFL students in the English department, and gave retrospective interviews. They were also asked to respond about anything in the instruments and procedures that they found to be difficult or confusing. Each teacher attended the session individually.

3.7.2 Benefits of the Pilot

The pilot study served several objectives at different levels, personally, academically, and with respect to the fieldwork practicalities.

Personally, it gave me the chance to develop my ability to conduct TA and retrospective interviews, to pose relevant questions that suited the context of investigation, and to

record the appropriate time. Moreover, it gave me the confidence that I need to approach the interviewees without feeling worried.

From an academic perspective, the Pilot allowed me to test suitability of the interview questions and instructions given. In addition, it allowed me to evaluate the TA training tasks.

From a practical fieldwork perspective, we were able to check the appropriateness of the room settings and the equipment. Regarding the informants, the pilot allowed us to evaluate the teachers' ability to understand and perform the think-aloud reporting and check for any procedural problems.

A number of important points emerged which affected the main study.

Firstly, the Pilot gave me valuable experience in controlling the time spent on the retrospective interviews and making sure that the teachers concentrated on the main questions.

Secondly, regarding the instructions, they were delivered to the teachers in two formats: recorded as part of the practising of the TA reporting, and printed in front of the teachers during the main TA task for those who need visual instructions. According to Bowles (2010), "pilot-testing helps to ensure that the verbalization instructions are written clearly, in a way that participants understand and can follow" (p.116). In this part, none of the teachers reported any question or difficulty in understanding the instructions. Yet, I perceived that I should ensure in the main study three importance points: a) the teachers should treat the rating task as a real task; b) they should think aloud and not direct their talk to me, they should imagine that they are alone (note these points are provided in the instructions but it seemed the teachers needed more emphasis); and c)

the participants' main focus should be on the rating task and not on thinking aloud (Dörnyei, 2007).

One more issue with instructions that arose during the pilot study was that there was no clear direction for the teachers on how to approach the rating task itself. For example, one teacher was asking me whether she should arrange the papers first or pick up the first paper at the top. Another teacher, asked me whether she should read the scripts out loud or not. Hence, I decided to make it clear in the main study that they are free in how to approach the task and should follow what they usually do in their real task situations.

Thirdly, for the think-aloud training there were initially two training/warm-up tasks as mentioned earlier. The first one was arithmetical and the second one was verbal. On the arithmetic task, the targeted teachers performed well but we realised that the task was rather remote from what they in fact being trained for. For the anagram, the second task, I used the word *disaster* which proved difficult to guess. None of the teachers could solve this task and they spent a longer time than I expected. Moreover, in the middle of the task, some of them lost the sense of how to think aloud properly. Therefore, I decided to resort in the main study to training them on the same sort of task which they would do for the study itself.

Fourthly, the pilot study confirmed the appropriateness of the room setting and the equipment. The teacher's and researcher's tables were arranged in an L shape. I sat on the long side of the letter and the teacher on the short side (see Figure 3-1below).

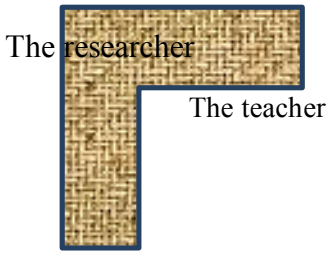

Figure 3-1 The room setting during think-aloud and retrospective interview

This setting had many merits. First, it allowed me as a researcher to observe the teacher interaction with the paper and note down any written comments from nearby without being noticed by the teacher and annoying her. This setting gives some kind of privacy to the teachers while they are thinking aloud. Furthermore, technically, this setting allowed me to place the recording equipment and the papers in one place from the beginning of the task without needing to move them.

Fifthly, the data in the think-aloud protocols crucially revealed that it was not possible to trace the raters' decision-making process for several reasons. Firstly, one of the writing tasks chosen was not an essay writing task as we intended. Instead the task was divided between gap filling, correcting wrong sentences, and putting right or wrong in front of sentences provided. Another problem was that, even where there was genuine writing, the students produced only short paragraphs, which were not enough for proper evaluation. While we could have found English majors of later years of study writing more suitable compositions, the problem then arose that they had only one writing teacher, whereas in our main study we aimed for ten. Hence, in the hope of obtaining more suitable data with enough available teachers, we changed the focus of the study to a different context, which is the UK ESL context.

Sixth, the use of a general interview appeared unproblematic, so we retained it without change in the main study.

3.8 Validity and reliability of the data collection

Undoubtedly major contributors to the validity and reliability of the data collection were the triangulation achieved by the use of more than one data source, and the Piloting, and consequent improvements, as just described.

Although the validity and the reliability of verbal reports have sometimes been questioned in the literature, advocates of these methods assert that well designed and conducted studies are reliable and valid (See Cohen, 2011; Ericsson, 2002; Gass & Mackey, 2000). Green (1998) highlighted two types of validity and reliability issues that are related to think aloud self-reporting. One concerns the validity and reliability of the data collection technique and the other concerns the validity and reliability of coding the data. Both of them can be sources of error and contribute to the success of the research. Furthermore, both apply to interviews as much as to TA. This section will discuss the validity of the data collection phase and leave the validity of the coding for 3.10 below.

3.8.1 Validity of the data gathering

Validity of the TA reporting refers to "the extent to which the verbalised information that is actually heeded as a task is being carried out corresponds closely with what is then verbalised, and the extent to which the task may be carried out without disruption by the requirement to generate a verbal report" (Green, 1998, p. 10). The issue of reactivity, which is that "the act of thinking aloud potentially triggering changes in learners' cognitive processes while performing the task" (Leow & Morgan-Short, 2004, p.35) has generated some investigation. In our study, we attempted to combat this issue by making as few demands as possible on participants in what we asked them to do during the TA. That is, we did not ask for detailed explanations of what they were doing or thinking.

Interviews may also suffer validity issues of a different sort. What emerges from less structured interviews may vary depending on the social skills and personality of participants, while in more structured interviews, cognitive ability of the participant

may have more impact on the data generated. Some interviewees may simply have a greater capacity than others to reflect on what they did in the recent task and express it overtly for the researcher. The researcher attempted to combat this by having some extra questions in reserve if teachers did not say much (e.g. just gave a 'yes' answer.

3.8.2 Reliability of the data gathering

Essentially, reliability of TA verbal reporting refers to "the likelihood that similar verbal reports might be produced by the same individual presented with the same or very similar tasks" (Green 1989, p. 11). We were in a way able to check this in our data since each teacher rated 6 or 7 scripts in succession. In general, we did not see any great script by script variation in how an individual teacher tackled the rating of each of his/her different scripts. That is, by and large, each teacher used the same criteria and followed much the same sequence.

Green (ibid) also mentions that individual differences and task variables are among the factors that affect reliability of verbal protocols, although they influence performance in fairly systematic ways. Contextual variables are another factor that influence reliability of TA reporting. Ericsson and Simon (1984) maintain that providing sufficient contextual information and minimising pauses (time lapse) between task and verbal report can boost the reliability of the data. This we achieved by having the participants report aloud concurrently with performing the task.

This last factor can also affect the reliability of retrospective interviews. In order to counter this, we held them immediately after the think-aloud for each composition and the context was maintained: the same room, and students' written compositions were kept in front of the participants.

3.9 Qualitative data transcription and analysis

Qualitative data emerged from open response items in the general interview, and from the TA and the immediate retrospective interviews.

The general interview open response data was not large and was transcribed and submitted to conventional content analysis to identify different types of point made in response to each open question. One such question asked teachers to add to the list we had provided any other aspects of text which they deemed to characterise 'good writing'. The responses were read over repeatedly, identifying and refining distinct categories of suggestion made. This was done with an eye to aspects of writing referred to in the literature as indicative of 'good writing', yielding the types that we report in chapter 4.2. Other areas coded were sources of rating criteria, students' problems, and rater raining.

The main challenge was handling the TA data, and its associated retrospective interview and observation data, which was far more extensive and occupied many months to handle and code, involving a number of false starts and changes of direction. Hence, we devote the rest of the account below to that. TA protocols related to the rating of 38 scripts were collected from the six raters participating in the study, since each rater produced verbal protocols of TA with retrospective interviews for six or seven written texts that were rated.

This section will first explain the procedures of how the data was organized and transcribed for the analysis. After that, I elaborate on the procedures used to segment and code the data. At each step, I present my approach to data analysis and the rationale for choosing it.

3.9.1 Initially preparing and organizing the data

The data was in two formats. The first was the audio recordings of the think-aloud protocols and of the retrospective interviews. The second format was the notes the researcher made of relevant non-verbal actions during observation of the TA, including teacher writing on the scripts.

In order to effectively use these types of data, they were organised in such a way as to reflect that the think-aloud protocol was regarded as the primary source of data for discovering the actual rating processes of the ESL teachers, whereas the other forms of data were needed in a supportive role to triangulate the findings from the think aloud protocols and provide deeper insights into the data.

Both the think aloud-protocols and the immediate retrospections were transcribed in succession in the same text file for each participant, later transferred into NVivo, in order to facilitate data retrieval, evaluation and decision making during the coding stage.

3.9.2 Transcribing the data

The aim of this step was to represent the data in a form that allowed us to handle it electronically. We followed clear transcription guidelines and created a good transcription sheet design, which accommodated data from our different sources (see example in appendix J).

An initial step in the process was to decide on clear transcription conventions as suggested in the literature (e.g. Gibson & Brown, 2009; Richards, 2003). This study used as a starting point for the transcription the conventions in Cumming et al. (2002) and Lumley (2005), as these two studies shared the same purpose of this study; this

helped me through the first few transcripts until I decided on adequate symbols to be used in the rest of transcriptions (see appendix I).

In their definition of transcription, Gibson & Brown (2009, p. 109) warn that transcription is not only "writing down what someone or some people said and did" but it involves critical judgment of how to represent and what to represent. Moreover, it should include choosing what to display and which features of speech, action or interaction should be in focus rather than others. Others such as White et al. (2007) however simply recommend the researcher to document as many features as possible of oral speech such as intonation, inflexion, pauses, variation in the rate of speech, sighing, swallowing, as well as accompanying non-verbal communication (such as pointing and gesturing); and physical actions (such as taking notes, turning pages, underlining). In our case, it should be noted that participants were audio-recorded not video-taped, since we felt that non-verbal information during the TA could be recorded sufficiently for our purposes from the researcher's observation notes, and inspection of the composition scripts.

We then set about doing the transcription using the advice of Richards (2003, p. 182): "The best way of deciding just what symbol to use is to get down to the business of transcribing in order to gradually develop your familiarity with the system and our sense of what needs to be captured in the talk." The reports were transcribed in the language they were produced in, which was all English. The decision was made to transcribe all protocols as heard and not to correct raters' mistakes, following principles established in earlier research (Lumley, 2000; 2002; 2005).

The transcription was made in English orthography with additional indication of pauses (indicated in seconds) and shorter hesitations. Following the recommendation of my

supervisor, I transcribed all the raters' talk. Utterances of the researcher during the TA had however been minimal (mostly in the form of *mm', 'mhm'* and *think aloud please*) and this feedback was omitted from the actual transcription. Utterances of the researcher during the interview phase were however transcribed.

Aware of the potential danger of losing data at the transcription stage, we were initially attentive when transcribing the TA to those sometimes neglected verbal (e.g., intonation, emphasis, tone) and nonverbal features (e.g., sighing, coughing, clearing throat, and noises) in the verbal protocols. Hence the TA transcription also included the researcher's observation which gave careful consideration to those spoken or non-spoken behaviors which may amongst other things signal teachers' (raters') affective reactions (e.g., groaning, exclamations, writing on the sheets, etc.) which can be helpful in the interpretation process of the data and to help the researcher to spot any similarities and differences between the teachers (raters) (White, Schramm & Chamot, 2007; Dörnyei, 2007). This required playing the voice recorder repeatedly and matching points in the recording with points in the researcher's observation notes in order not to miss such instances. Thus, after transcribing each section in the data, the information from the observation notes (which includes non-verbal communication and physical actions) was added in a separate column.

After doing this for two teachers, however, it seemed that the amount of time and effort involved was not matched by the benefit in information obtained that was relevant to our concerns. Hence, we discontinued transcribing the extra information and fell back on a simple orthographic transcription for all the remaining data. After all, the data had been recorded in order to analyse the content of what was said, not primarily the way in which it was said. Hence a highly detailed phonetic and paralinguistic transcription was not in the end deemed to be needed.

The protocols were all lengthy with the TA components varying between approximately 10811 words and 18595 words as Table 3.3 shows. Overall, the transcription stage took a few months to complete.

Table 3.3 word count for TA components of data[4]

Rater	Number of scripts	Approx. word count
T1 James	5	8632
T2 Bob	5	10811
T3 Zain	7	17,092
T4 Sam	9	18595
T5 Gena	7	11710
T6 John	6	13239

The transcription sheet for each teacher was designed to hold a record of the teachers' demographic information from the general interview, and to accommodate my notes as well as the transcription of what was recorded, the think-aloud protocols and immediate retrospective interviews (see Appendix J for an example of the transcription sheet design). It consisted of two parts. The first one was dedicated to recording the participants' basic details, such as ID number, the session booking information, the whole session recording time, and the start and finish time for the TA of each script. The second part of the transcription sheet was dedicated to the actual transcription. This was designed in the form of a table with six columns. The first column was for recording the stage of the task, e.g., the rating task with TA, or interview retrospection part and the number of the script. The second column was for time; the third was for speaker; the fourth, for the transcription of the talk itself. The fifth column was for my observation notes about each teacher behaviour and non-verbal actions next to the

[4] These word counts do not include researcher feedback

transcription of the relevant part of the think-aloud protocol. The last column was for my notes during the transcription stage to enter my preliminary analysis of each teacher's behaviour that would help me during the coding stage. The last column however did not determine the data analysis later on (see Appendix J for an example of a completed transcription sheet).

Overall, the transcription stage took a few months to complete. Following this we worked on segmentation and tentative coding with the data in Word files on computer.

3.9.3 Approach to segmentation and coding

Before discussing the ensuing analysis procedures, we need first to introduce the approach to coding and the rationale for choosing it. The approach of the present study was largely inductive, following some principles of grounded theory, and partly deductive, exploiting categories known in advance from the literature. The codes emerged from two sources: from the data, itself and from a list of a general codes that I prepared from other studies beforehand. Simple adoption of coding schemes used by earlier researchers including Huot (1988), and Cumming (1990) was considered but in the end found to be unfeasible. This was parallel with the view of various researchers (e.g., Huot, 1988; Smagorinsky, 1994a; Green, 1997) that the categories need to be developed to fit the data gathered in each specific context. Smagorinsky (1994a) states:

> "the work is very "messy" indeed and […] researchers often have little precedent to follow in conducting particular studies. Even clear guidelines are often of little help during the conduct of unique investigations." (1994a, p.x)

Moreover, Smagorisky (1994c) argues that importing a coding system from other studies is no help as it was designed to answer other questions. Nevertheless, in our study an initial list of codes was found useful, based on previous studies (Vaughan,

1991; Erdosy, 2000; Saki, 2001; Cumming et al. 2002; Lumley, 2005).

3.9.4 Initial exploration of part of the TA data

The notes which I took while observing the TA helped me a lot to engage initially with the data and gave me at least a preliminary judgment of how to analyse it. Moreover, as described above, during the transcribing stage and on the right side of the transcription sheet I wrote my observation notes and I began to add my analytical notes and jotted down a tentative list of preliminary codes. All this helped me a lot to immerse myself deeply and provided me with a good sense of the data which guided me later.

At this point a decision was made to print out two participants' transcripts and attempt a complete analysis of these in order to assist in developing a system which could then be applied to all the data. This step gave me the freedom to write and correct myself all the way through and helped me in going forward and backward through the sheets more quickly.

It had emerged impressionistically during the transcription stage that participants with different educational and experience backgrounds used different assessment criteria. Therefore, the complete database records were printed out of two participants (T1 & T3) who differed notably in these respects to see to assist in developing an initial feel for segmentation and coding which would deal with the full range of different phenomena that might occur in the data.

We then began reading the data, highlighting or underlining the evaluation criteria that emerged from these teachers while they were reading their students' compositions. As I mentioned earlier, there was no rating criteria or scale provided for the teachers to

refer to, so this was a great challenge for me. Consequently, several issues came to the surface at this point which informed the later analysis.

The first issue was related to how to collect and identify the criteria employed by a single teacher who evaluates many scripts and how to decide if he/she used the same criteria all the way through. In order to track this, I read all the TA data produced by one teacher and highlighted all the criteria he mentioned, including repeated criteria. From this I obtained a general idea about this participant and his evaluation stance. This triggered another issue which was how to decide if a certain criterion was a frequent element in all his/her behaviour and not an isolated item that was used just once. The best way to decide was to look at the retrospective interview and my analytical notes to see whether he/she emphasised this criterion or not, and also compare it to his actual behaviour in all the think-aloud transcriptions. This process had later to be applied to all informants.

Another important issue was how to formulate different and specific codes for all the different participants' criteria and develop a common coding that fitted all the data. To do this, I later had to go back and read again all the transcriptions for all informants and highlight the same criteria with the same colour and check them against their immediate retrospections. Then I had to check my analytical notes and find any evidence of a match with the teachers' behaviour and highlight that as well. Then, I coded it using descriptive codes.

Another issue was how to know what should be coded from each data set. As we mentioned, we have three sources of data that describes the same event, so dealing with more than one source can be misleading: if we code the same behaviour several times this may exaggerate the frequency of the behaviour coded. However, as I mentioned

earlier, the think aloud instrument is the main instrument that leads this research and it is the most valid instrument to find out about the rating process, hence eventually I mainly coded the criteria from the think-aloud and used the data from other sources to confirm and illuminate my coding decision.

The fourth issue that emerged was how to deal with the discrepancies between evidence from different sources.

Finally, although at this point our focus of attention was on coding criteria, it became apparent that there were many kinds of thing said in the TA which were not just references to criteria, and which would require greater attention later.

Having gone through the two printed protocols and familiarised myself with the issues which emerged, I then started the segmentation process.

3.9.5 The preliminary segmentation of all the TA data prior to full coding

The next stage we felt to be needed prior to a full coding was careful segmentation of all the TA protocols collected. Segmentation is performed prior to coding as way of identifying those stretches of transcript which signal something that is felt to require a separate code, even if at this point what the code will be is not yet clear. It is a controversial matter, which some researchers define as arbitrary and intuitive (Shirazi, 2012; Lumley, 2005). In qualitative research, the segmentation process often involves breaking verbal data into separate chunks or idea units of different lengths (e.g., phrases, clauses, sentences, etc.) to be later categorized and coded. Even at this stage certain tentative labels may be assigned to the segments (Dörnyei, 2007).

Studies have varied in types of segmentation procedures applied in analysing think-

aloud protocols. Ericsson and Simon (1993) point out that in speech the boundaries of phrases are usually signalled by the pauses. Linguistic boundaries cannot always solve the segmentation problem, however. Paltridge (1994) also suggests that segment boundaries in transcripts need to be decided on the basis of the content of the texts instead of the way in which the content is expressed linguistically. Moreover, he claims that the divisions within texts are 'intuitive' and frequently made in the absence of "textual indicators" (1994, p.295).

Someren, Bernard, and Sandberg (1994) put forward another method for segmentation. They suggest that "the combination of these pauses and linguistic structure provide a natural and general method to segment a think-aloud protocol" (1994, p. 120). They underscore that a high level of agreement between people exists while they listen to think aloud protocols. They believe however that what makes segmentation more difficult and less reliable is to segment the protocol on the basis of the written transcript only. Green (1997) by contrast pays more attention to the content, and suggests that each segment should represent a different process and will usually comprise a phrase, clause or sentence. Lumley (2005) on the other hand, believes that this process is simply arbitrary as it leaves "a lot of scope for decision making" (p. 135).

Taking account of the above views, eventually, it was decided that in our study segmentation into segments or text units (TUs) would be guided by attention to a combination of linguistic structures and pauses plus a small number of other principles. Above all, however, the segmentation of the raters' think aloud talk was based on units of thought or content. In addition, the following guidelines were developed for segmenting the TA transcripts.

- All the talk associated with beginning the rating task and identifying the script was considered as a single unit because it was dealing with instructions that had been given to the raters by the researcher about what to say at this point. A separate text unit was assigned for identifying the name of the student if mentioned.
- A separate segment / unit was used for:
    - Each piece of text read aloud from script
    - Each separate comment made by the rater. This raised several dilemmas since some comments could be perceived as concealing single or multiple ideas and sometimes it was impossible to determine when a rater begins to express a new idea. However, the appropriate approach was to attempt to create a new text unit for each new idea or topic or apparent change of direction. The idea was to create too many texts units rather than too few.
    - The nomination of an evaluation category was considered a separate segment.
- The rater's first reading of the essay was considered as a single unit even if disrupted.
- When there was repetition of the same idea or phrase, it was treated as a single segment.
- When the rater identified many evaluation features in one sentence (for instance, spelling, grammatical mistake, punctuation) each one of these was considered as a single segment.
- Any comment about giving the score in the final evaluation was treated as separate.

- Reading part of the script, the whole script, or rereading the script were each treated as one segment.

Excerpt 1 presents an example of segmentation arrived at this stage for the start of a TA protocol. The rater first identifies the script number, then mentions the essay topic and thirdly reads part of the title. He next reads the writer's name (4), followed by activating background knowledge of the writer, then (6) indicates that he will start reading the text. Excerpt 1 *Example of segmentation of TA at the start of a script*

RZ1. Script N5

TU Teacher's Talk
1. This is umm number 4, 5. It is 5.
2. so the topic was about child work.
3. Do you think that children can work and get paid um…
4. Who is that? This is Barbra from Turkey.
5. Umm, as far as I know she got problem in grammar I think.
6. So let me start.

In Excerpt 2 we show an example of segmentation in part of the data where criteria and scoring were covered. The rater offers hedged positive evaluation of grammar (80) with the words 'has potential' and states his negative evaluation of one specific grammatical feature (81) and how it affects the score for grammar, coincidentally implying that he is working with separate rating scales for different aspects of the language (82). Another feature 'Vocab' is targeted (83) with the positive word 'OK' hedged with a negative reference to being 'not very advanced'. 84 shows the rater's negative evaluation of 'spelling mistakes' in the text. In 85, there seem to be reference to a previous negative

judgment about ‘paragraphing’, followed by deliberation on the score given in 86, before making a final score judgment (87).

Excerpt 2 *Example of TA segmentation containing both positive and negative evaluation on various criteria*

R3J. Script N2

TU Teacher’s Talk
80. As I said, the grammar has potential,
81. but because of the uncountable nouns problem
82. she’s kind of harmed the score for that.
83. Vocab is OK but not very advanced,
84. and there are spelling mistakes, which is a pity.
85. But yes, the paragraphing is the big problem.
86. Yes, I’d say overall Grazia gets a 5. Not so far from a 5.5,
87. but it’s a 5 at the moment.
88. So that was number two.

Transcript segmentation raised several dilemmas. One was incompleteness of some comments. Transcripts are recording what was said rather than thought processes directly, and raters focused on the rating task not on completeness or explicitness of utterances. That is why segmentation followed the principle of content above linguistic features. However, some comments were left unfinished or contain opaque references. In Excerpt 3 we see three examples of this. In 11, T4 fails to complete saying what he is going to 'make'. Later (12) he says that that the student 'is using this’ but the reference of 'this' is not clear to the researcher. Again in 13 the rater referred with demonstratives to a previous topic in class teaching without saying what it was. Since segments are meant to be potentially codable chunks, these all were problematic since the lack of information made it hard to see how they could later be coded.

Another issue was repetition. In Excerpt 12-15 are multiple utterances about the same thing, the voice of a specific bit of student text. Some are similar in wording, others less so. This then raises the segmentation issue as to whether these are all the same segment or not.

Excerpt 3 *Examples of incomplete or inexplicit comments*

TU teacher's talk
11. I am going to make a …. for extra marks
''He is fairly clever''
12. um... he is using this….
13. I was telling him that in class. He did not seem to know this! But he does!
14. that is 10 and 11.

Excerpt 4 *Example of repetition of the same idea*

RZ1. Script N2

TU Teacher's Talk
…should be childhood…
12. So again article. "childhood should not have burdened… uhmm…
13. "should not have burdened by x hours." Should not have burdened." It should be: should not be burdened.
14. So I would say… um um: "should not have burdened,''
15. So I would say; active passive. So she has to be attention to the active passive voice.

These working analyses, as produced at this stage, showed that the process of segmentation of the text and a preliminary coding analysis are closely intertwined. The segmentation process is itself an iterative process subject to continual modification in the course of the data analysis (Miles & Huberman 1994; Green 1997).

To sum up, transcripts were tentatively segmented into TUs based on content; TUs were numbered. At this point we additionally provisionally identified the following types of content in the units that emerged, ranging much wider than just different types of criteria:

- Identification of each script by mentioning the number or name.

  e.g. Let's start with number 1, she is a Saudi lady from C1B.

- Reference to reading as part of the rating process

  e.g. I am going to read this again to make sure.

- Criteria that raters refer to on the way to making their decisions

  e.g. Usually, these essays are full of grammar mistakes,

  e.g. That is a nice introduction.

  e.g. While the first student was so grammatically accurate actually. I focused mainly on ideas

- Reference to the text and/or the writer

  e.g. She is so intelligent, she knows how to write complex sentences and she is supporting her ideas.

- Non-verbal act

  e.g Will put a tick.

- Justifications made for specific rating or score

  e.g. And if she provided another paragraph before the conclusion, she would get 8, perhaps

- Activation of knowledge of the student background or level

  e.g. This is (Ba**) from Turkey.

  e.g. Umm, as far as I know she got problem in grammar I think. So, let me start.

  e.g. Ok who is this? Oh, this is (**za). She is nice. She got 7.5 in listening. So, let's see how she writes?

- Problems raters encounter in the rating process.

  e.g It is difficult to read his handwriting.

3.9.6 Transferring the data into qualitative data analysis software

Having spent considerable time on the above procedures, there came a point where it was felt that it was necessary to find a better medium to store, manage and retrieve the data which would also assist with the full process of coding. For this purpose, the most popular qualitative research analysis software (NVivo 10: Blaney J., Filer K., & Lyon J., 2014) was utilised. Aside from storing the data in files and folders and allowing it to be sorted in a variety of ways, it assisted the researcher in developing her coding system to capture the themes she identified in teachers' evaluation of students' essays. Data analysis is otherwise more time-consuming and likely to miss things when analysing data in Word files, for example. In addition, it helps in presenting findings.

NVivo 10 was chosen for several reasons. Firstly, as mentioned earlier, it assists the researcher in manipulating data, browsing it, coding it, and annotating and gaining access to data quickly and accurately (Richards, 1999). Secondly, NVivo has tools for "recording and linking ideas in many ways, and for searching and exploring the patterns of data and ideas. The various actions - proxy documents, nodes, and attributes give values to attributes, linking, coding, and shaping in sets the documents and nodes" (Azeem & Salfi, 2012). Thirdly, it is supported and provided by the University of Essex.

To begin with, all the transcribed files were uploaded into the programme. Then a person node classification sheet was created to categorise and arrange the participants' demographic data (See Appendix J). Once all the data files and the participants'

demographic information were set up in NVivo10, we were ready to work on a full analysis of the data, using what we had learnt up to this point.

3.9.7 Starting point for the full coding scheme

We now move to the full TA coding stages, recognising that many ideas for it, as well as a provisional segmentation, had already arisen from the work reported above. Through much of the work from this point on we felt torn by several opposing forces.

On the one hand, the fact that raters in this research were given no rating scale to follow and every rater was left to his/her own criteria and scale for rating meant that finding common codes that fitted all data was quite a difficult task. We were also aware that research based on TA verbal protocol analysis has been said to have one feature in common, that there is no single uniform coding scheme and some say a specific one should be developed for each study (Ericsson & Simon, 1993; Green, 1997).

On the other hand, we found that there were common areas between raters in previous research and raters in this study that meant that codes could be shared or adapted. This presented the attraction of drawing on existing taxonomies in the literature. At one point, we were attracted by the simple division of all codes into three high level categories of behaviours that could be used as a starting point for coding scheme development for the present study:

- management behaviour
- reading behaviour
- rating behaviour

This however did not in the long run prove useful. At various points, we drew on coding schemes of Cumming et al. (2002), Lumley (2005), and Barkaoui (2008).

At the same time, we were aware that it was necessary to arrive at a final coding scheme that would describe the data in a way that would help answer our research questions. We needed to unpack broad terms like 'marking practices' or 'decision making processes or 'reading behaviours' which had been met in the literature and focus on capturing insights to illuminate an emerging set of topics which ended up somewhat like this:

- criteria that raters refer to while evaluating, and how they were weighted relative to each other
- how raters obtained and justified an overall rating or mark from their detailed comments
- What other kinds of comments teachers made, beyond those related strictly to the rating itself
- What common sequences of activity appeared in the rating behaviour of teachers

What distinct rating styles appeared in different teachers' Full TA coding went through several stages as described next. As can be seen from this account, progress was not linear but characterised by attempts to use a variety of different schemes and focus on different aims at different stages leading to a tortuous voyage of discovery.

3.9.8 The first cycle of full TA and retrospective interview coding

I started the first coding using Nvivo 10, based on my preliminary list of codes from other sources and on my previous experience with the earlier two printed think-aloud sets of data and with the segmentation process. The preliminary list of codes was organised under six main categories.

The following step was an attempt to use this list of codes to code the data and to see how to best use NVivo10. In this stage, I first started by using the two printed sets of data which I had coded manually. Using this data made it easy to figure out how well the main categories worked. In addition, they helped me to deal with NVivo10 without too much difficulty. For this step, in NVivo10, I chose to create structured memos to record the non-verbal data and my analytical notes. These memos helped me to link the verbal protocol transcript with my observation and my analytical notes as well as to use the content of these memos to validate and support the coding of the think-aloud transcript (Wiredu, 2014).

At this point I moved on to code the rest of the TA data. The purpose of this step was to generate more codes and start refining and defining the existing ones. These were ongoing processes. I was flexible in adding new codes under the main categories as well as adding main categories when needed, and made necessary modifications to their names or definitions according the nature of the data.

The first cycle of coding involved a lot of further decision making. The type of decisions I made are illustrated in Table 3-4 below.

Table 3-4 Examples of decisions made during the first cycle of full coding

	The issue	The coding decision made	Rationale
1.	When the teacher reads aloud and then s/he stops for evaluation or for interpretation and then s/he does the same for the	Code as: *Read part of the text*.	The reading behaviour is an essential foundation of the rating process. Hence, we shall

	whole composition, reads then stops.		differentiate between reading stages.
2.	When the teacher spots an error in the Grammar category and just articulates it, e.g. *this is a grammatical error*, without any evaluative comment.	Code as: *Classify error into types*.	
3.	How many times to code *Read part of the text* when the teacher's behavior is to read then stop?	Code it once if the rater stops only to give minimal feedback such as *ok/good* or articulate an error and then move on. However, if the rater stops to give detailed evaluation, I will code it a second time after that.	Since we planned to count code frequencies in the end, this seemed to be the best compromise between obtaining artificially high numbers of this code and unrealistically low ones.
4.	Some teachers read all the text at the beginning but it is not clear whether they are reading all of it or just scanning it.	Two different codes: *Read whole text*; *Scan the text*.	As I mentioned earlier reading behaviour is an integral part of the rating process, so I have to be more specific in this code. Scanning could be identified

			through teacher's verbal comments.
5.	How many times to code *Reread*	If there is no interval between *Reread* and doing something else (such as *Reading the task prompt*) code it once. However, if there is an interval between rereadings then code it twice. If the teacher rereads for the first time and then tries to interpret the sentence to understand it and then rereads the same sentence or same paragraph for more clarification or evaluation then code it a second time. Moreover, when the teacher missed something out and recalled it and then reread the same sentence/ paragraph for more evaluative purposes, code as another rereading.	The word *Reread* itself implies that the material is read for the second or more time. Hence, regardless of the number of times the teacher rereads the text it should be coded once unless something substantial intervenes

6.	When something is said that does not refer to teacher evaluation or actual decision making, such as teacher saying 'I will put a tick or arrow' or 'I will underline this' without classifying what the kind of error is.	I will be coded it as *Teacher's reaction.*	We do not need to get into too much detail on anything written or said that is not informing us about rating criteria and decision making.
7.	When the teacher writes on the script some comments or abbreviation that denotes error types such as SP or VF.	Not coded	Because this concerns feedback to the writer, which is not the target of the study.
8.	Both in the TA for the researcher and on texts as feedback for students, teachers often indicate either positive or negative evaluation such as: 'the organization is good' or 'there is no organization'. Shall I set up two different sub-codes -	No. all these should be arranged under one single code for the aspect of writing/criterion that is in focus (e.g. organisation).	This decision will eliminate the chances of discussing feedback on writing which is a different angle from my research. We are concerned with what criteria are used in rating, how weighted, etc., not in whether the

	positive and negative codes - under every criterion?	essays are often rated positively or negatively on any criterion.

In coding teachers' decision making during their rating process, I felt it was important to distinguish between where they give immediate responses to a specific part of the text and overall comment about the whole text, avoiding going deeper. Hence, I introduced an independent code for *Overall impression.*

The result of the first cycle of coding was a list of 55 codes under five main categories: *I. Focus*; *II. Presentation; III. Task requirements; IV. Reading behaviour* and *V. Non-evaluative comments.* This cycle produced the initial list of codes that constituted a useful start to code all the TA data and helped me to gain an overall picture of what was going on during the think-aloud process. However, these codes needed to be taken under careful review and evaluation in the next cycle of coding.

Since the main categories and rationale of the coding system and the list of specific codes later changed considerably from the one sketched here we will not describe it in more detail.

3.9.9 Reviewing the list of codes after the first cycle of coding

The list of codes that emerged during the first cycle was reviewed and examined with the help of my two advisors and with an expert in qualitative research software to increase the validity of the analysis. The process of code examination encompassed many aspects. Firstly, it looked at the content of the codes, i.e., what they cover or reflect. It also looked at the length of the code list. Furthermore, it covered the names of the categories and the relationship of the code to each category. The following

paragraphs discuss each area of the code examination and the decision made in relation to them.

Regarding the first area of examination, which is the content of the codes, the review showed that the codes cover many different content areas, most of which are evidenced from more than one teacher. The codes were designed to fit all the data produced by teachers' think-aloud and relevant parts of what they said in the immediate retrospective interviews. The system includes codes for rater (teacher) actions that are not in themselves evaluative: for instance, the code *Reading behaviour* is a general code that covered five sub-codes that represent individual behaviours, e.g., *Read the whole text*, *Reread*, *Read part of the text, Scan whole composition* and finally *Read essay prompt*. Furthermore, there were many other codes that represent forms of evaluation, e.g., *Consider error gravity* and *Consider error frequency*. Moreover, there were subdivisions of certain main codes to capture further aspects of that code, as seen for *Reading behaviour* above and again for *Classify error into types* which included many sub-codes like: *Verb form, Sing./pl. noun, Prepositions, Plurals* and *Articles.* In the analysis, I tried to concentrate mainly on the evaluative writing criteria without looking too much at how teachers give written feedback.

The second area of concern was the length of the code list. A close analysis of the code list however showed that its length needed to be considerable for several reasons. Firstly, the nature of the present study obliged me to design a considerable number of codes to fit the different behaviours of teachers as each one had his/her own procedure and criteria in assessing their own students' writing. Indeed, in the final coding, 59% of the codes occur in data from only two or one rater and none occur in data from all six teachers. Furthermore, in our analysis it is inevitable to describe reading behaviour as an essential component of writing assessment. Hence, there had to be detailed codes

for *Reading behaviour* as we mentioned earlier, which inflated the number of codes. Moreover, no one can deny that the *Language* aspect is an integral feature of any writing rating scale. Hence, our scheme of codes needed to cover this aspect with its specific criteria, e.g., *Grammar*, *Mechanics* (spelling, punctuation, and capitalisation), *Layout* and *Language fluency*. One more reason is because, due to the special interest of our study that investigates teachers' rating of their own students' writing, I created a code for *Teachers' reaction* to accord with the uniqueness of this research. Hence it was of importance to my research to keep many codes regardless of their effect on the length of the code list.

Furthermore, the review considered the names of the categories and the grouping of the codes under each higher category. For example, the category *Cohesion and coherence* seemed initially to refer to a general aspect of writing assessment but we later realised should be part of the *Organization* category. The *Vocabulary* code was also placed under the *Language* category, and so on. These issues were considerable, however, and the final code categorisation was only arrived at after the second cycle of coding.

Finally, the review helped me to refine some codes and exclude others. For example, there were a *Positive evaluation and a Negative evaluation* codes that ware excluded from the coding list.

Having reviewed the codes, the next step was to apply the outcome of this review to the list of codes we had. First, I created a new folder in Nvivo 10 called the Thematic Coding Framework in which I copied the first list of codes from the previous nodes folder. In this new folder, I copied the refined codes only and left out the unrefined ones. Next, I added some new codes that captured in more detail teacher's reactions

which were not evaluative, for example: *Classroom teaching actions* and *Emotional reaction*.

At this stage, it seemed reasonable to compare the list of codes I developed with other codes for the assessment decision making process in the literature to see if I missed something important. Hence, I looked at two list of codes from Cumming et al., (2002) and Lumley (2005). When I checked my list against these two lists, there were some common areas between them. The complete new list of codes at this point consisted of 72 codes.

The review phase required us to think deeply about the entire code list and therefore increased our general understanding of the codes. The outcome of this examination was used to start our second cycle of coding from a more informed position.

3.9.10 The second full cycle of coding and its review

The second cycle of coding involved using the reviewed list of codes, which emerged as an outcome of the first cycle of coding and its review, to recode the entire set of data in NVivo. The aim of recoding the whole set of data using this list was to ensure the validity and comprehensiveness of the list. According to Saldaña (2013, p.207) the primary goal during the second cycle of coding is "to develop a sense of categorical, thematic, conceptual and/or theoretical organisation from your array of first cycle codes". It is a further step in refining and finalising the list of codes.

After the second cycle was completed, the code list was reviewed again. This included examining names of the codes, their spellings and if there was any repetition and/or any overlap among them in respect of how each was defined. The final scheme converged to a considerable extent with taxonomies in other research on rating behaviour such as

Cumming et al. (2001; 2002), Lumley (2005) and Barkaoui[5] (2008). We also checked the frequency of some codes, excluding low level ones that were there from the first cycle of codes but not used in the second cycle. Above all, we arrived at a new high level organisation of the codes, in the end essentially adopting that of Cumming as more suitable than any we otherwise could find or think of (see 2.3.1).

3.9.11 The final list of codes

The final code list (Table 3.4) contains 64 rows, but some of these are high level categories like *Mechanics*, which do not occur separately: only their subcategories occur. Thus, there are 58 specific categories which occur. While many of the lowest level coded categories (= nodes in NVivo) are the same as in earlier versions of the coding system, the high level categories are rather different. Essentially all the TA data is now regarded as reflecting thought or action falling in one of three areas of Focus, and within each area the same binary division applies between what is called Judgment, and Interpretation, following Cumming et al. (2001; 2002).

Cumming et al.'s (2001;2002) overall scheme divides rater reported behaviours into three basic areas: 'Rhetorical and ideational focus' concerning rater references to writer use of rhetorical structure and content or ideas; 'Language focus' showing raters' attention to lower level aspects of the language like vocabulary and grammar; and 'Self-monitoring focus' which has a less clearly unified identity and really contains everything else that the rater refers to doing/thinking that is not covered by the first two, better defined, categories. In particular, it covers important areas such as the rater's mode of use of information sources such as the essay being read and background knowledge of the student, and the reasoning that the rater engages in to obtain an overall

[5] Barkaoui's (2008) coding framework was based mainly from Cumming et al.'s (2001, 2002).

rating from the detailed criteria considered.

Each of the three main categories is then subdivided into a 'Judgment' and an 'Interpretation' section. Again, this distinction is more transparent with the first two top level categories, where the Judgment strategies fit well with our central interest in the criteria used by raters. Judgment strategies concern the raters' evaluation of the scripts, whereas the interpretation strategies focus on the strategies which raters employ to comprehend the scripts or respond in non-evaluative ways.

The final codes are listed in Table 3.4, and a definition and example of each is provided in (appendices L and M).

One area, Language Focus, concerns the language of the compositions, in the sense of grammar, vocabulary, punctuation etc. Any evaluation (aka rating, assessing) of such aspects is coded under Judgment while other talk, e.g. just identifying something, is coded as Interpretation. The second main area, termed Rhetorical and Ideational Focus, concerns the higher levels of writing in the compositions, i.e. this covers what we previously called organization, coherence, topic and the like. Again, where any of these are evaluated it is coded as Judgment, otherwise as Interpretation. The third area, Self-Monitoring Focus, covers anything that the rater does or thinks that does not involve the script-based features covered in the first two categories. Under Interpretation it covers non-evaluative actions like reading the script or referring to student background characteristics, while Judgment covers some of the core areas of rating, such as the rater defining the criteria to be used or providing and justifying a score.

In conclusion, it can be said that the coding scheme evolved from a combination of (a) the evidence in the data, (b) input from advisers, (c) consideration of the coding schemes of Cumming et al. (2002), Lumley (2005), and Barkaoui (2008). This was

processed and synthesised over a long period of revision. I believe it combines the benefits of a top down (positivist) and a bottom up (constructivist) approach. The coding system underlies much of what is reported in the next chapter. In particular, it is directly reflected in the contents of Tables such as 4-6, 4-7, 4-11, 4-14 and the account based on that. It is not used as the sole basis for reporting the results since many of the research questions needed answers using a mix of the general interview, TA, and retrospective interview material, and qualitative attention to specific utterances at an even finer level than that which the coding system afforded.

Table 3.4 Final list of TA and retrospective interview codes

Code name

I. Self-monitoring focus

A. Judgment strategies

a. Summarise, distinguish or tally judgments collectively

b. main cause for given score

c. Ignore the error

d. Define or revise own criteria

e. Decide on macro strategies for reading and rating

f. Consider own personal response or biases

g. Compare with other compositions or students

h. Articulate or revise scoring

i. Articulate general impression

B. Interpretation strategies

a. Teacher attitude

01. Refer to teaching

02. Teacher's expectation

03. Provision for feedback

04. Emotional reaction

05. Other comment

b. Reflection on student's background

c. Identify student's name

d. Referring to personal situation of the writer

e. Reading behaviour

01. Scan whole composition

02. reread

03. read whole text

04. read part of the text

05. read or interpret easy prompt

II. Rhetorical and ideational focus

A. Judgment strategies

a. Task fulfilment and requirement

01. Assess relevance

02. Assess task completion

b. Relevance, quality, appropriacy of argument

c. Identify redundancy

d. Assess organization

01. Paragraphing

02. Main body

03. Introduction

04. Conclusion

05. Cohesion & coherence

06. Assess reasoning, logic, or topic development

07. Overall organization

B. Interpretation strategies

a. Personal reaction to ideas

b. Interpret ambiguous or unclear phrases

c. Discern rhetorical structure

III. Language focus

A. Judgment strategies

a. Vocabulary

01. Choice, range, appropriacy

02. Errors

03. General comments

b. Mechanics

01. Spelling

02. Punctuation

03. Capitalisation

c. Language fluency

d. Consider syntax or morphology

e. Consider error gravity or severity

f. Consider error frequency

g. Clarity, confusing, comprehensibility

h. Assess quantity of written production

i. Rate language overall

B. Interpretation strategies

a. Observe layout

b. Interpret or edit phrases

c. Classify errors into types

01. Word order, wrong word

02. Verb form

03. Sing. noun, pronoun

04. Preposition

05. Plurals

06. Articles

3.10 Quantitative analysis

Simple quantitative data emerged in several ways from our study.

First, in the general interview, there was a set of closed rating scale questions about what the participants thought were the features of 'good writing'. Responses were tallied.

Second, we counted occurrences of all the codes in the final coding system, for each person rating each script, so as to be able to talk about frequency of different rating behaviours.

Third we noted total words of TA transcript for each teacher, so as to be able to comment on each teacher's mean number of words of TA per script, and how that related to their background features.

3.11 Validity and reliability of the data analysis

Validity of the TA data analysis was mainly supported by involvement of experts, my supervisor and the advisers on my supervisory board, who gave valuable feedback at

several stages in the process, for example as described in 3.9.9. We also tried to ensure validity by taking due account of published coding systems in the literature, and indeed our final system was quite indebted to Cumming et al. (2002), who is considered a leading figure in writing research. Reliability was checked mainly in the form of intra-coder reliability. As shown repeatedly in the account in 3.9, the researcher went over and over the data an inordinate amount of times coding, recoding and checking coding. This we feel has resulted in a high level of intra-coder agreement, which supports the reliability of the coding.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Introduction

The aim of this chapter is to present the results and discussion of this study, which are organized so as to answer each of our research questions in turn. As we described in 3.2.3, we have nine research questions, some of which were answered primarily from the general interview data, not related directly to the actual rating tasks which participants undertook, and others which were answered mainly using the TA and retrospective interview data, based on the specific rating tasks performed by teachers for the study.

We begin with three topics which concern the teachers' general reports and views concerning key aspects of the assessment of writing: what general traits they believe characterise 'good writing' by their students (4.2), what rating scales and criteria they claim to typically use when rating writing, and where they come from (4.3), and what relevant training which they had received and how they perceive it (4.4).

The remainder of the research questions concern what they actually do when rating essays, based on their think aloud and immediately retrospective reports. We consider the criteria and weightings they reported while actually assessing the quality of the essays they were asked to rate (4.5), and the kinds of justification they offer for their judgments, beyond what was actually in the essay scripts which they were rating (4.6). We next review the other kinds of comments they made, not directly affecting their rating, and what this tells us about the nature of the rating task in this study compared

with others (4.7). In the next section (4.8), we consider the broad pattern of sequences of activity that the raters seem to follow and finally we attempt to identify the different rating styles that are represented in the teachers (4.9).

4.2 Teachers' general perception of 'good writing'

This section is designed to answer the first research question **RQ 1**:

What do our English writing teachers perceive as good writing?

Recall that the teachers were asked about 'good writing' in the context of writing done by their own students, and the data to answer this question was obtained mainly from the semi-structured interviews that preceded the think aloud rating tasks (See 3.5.2)

Part of the interview consisted of seven closed rating scale questions about different aspects which might contribute to 'good writing'. A summary of the teachers' responses is displayed in **Table 4-1** below. The results show that teachers mostly perceived all the features as highly important. However, when considering the rank order of importance, the two writing features perceived to be most important in good writing were task completion and appropriate vocabulary use, voted unanimously as very important. Next in overall importance came jointly cohesion, organisation and development of ideas. These were followed by grammar and finally mechanics. It is notable that the only low rated feature, by two teachers, was mechanics.

Table 4-1Teacher ratings of seven aspects of 'good writing'

Writing features	Not important (%)	Not very important %	Neither important nor unimportant %	Fairly important (%)	Very important (%)
Cohesion: unity of ideas and structure	0	0	0	50	50
Task completion	0	0	0	0	100
Relevant development of ideas	0	0	0	50	50
Appropriate grammar	0	0	0	66.6	33.3
Appropriate vocabulary use	0	0	0	0	100
Organization	0	0	0	50	50
Mechanics (e.g. spelling, punctuation, capitalisation)	0	33.3	0	0	66.6

These findings demonstrate that, in the teachers' view, the importance of writing features at the content and discourse level outweighed that of features at the linguistic level, with the notable exception of vocabulary. This partially echoes Cumming et al.'s (2002) outcomes regarding their ESL raters' views about the qualities of effective writing. They state "[the text qualities] they most frequently mentioned were (a) rhetorical organization, including introductory statements, development, cohesion and fulfilment of the writing task; (b) expression of ideas including logic, argumentation, clarity, uniqueness and supporting points" (Cumming et al., ibid; p. 72).

More importantly, apart from the above predefined features, some teachers in open response added other features to the above list which they considered as important components of good writing. These features were classified into two categories (3.9). These we called product and process, as outlined in Table 4-2 below.

The product-oriented statements, which are far more numerous, relate to writing features represented in the students' written texts, just like the seven features we

covered above, so are readily available to anyone assessing the text. Strictly, of course, they are not 'in' the text, but discovered through the reader's judgment of the written text through reading it.

The notion of the product-oriented approach to writing assessment is characterised by focusing on the content and organisation of the text, and language features (Cumming & Riazi, 2000). By contrast, process-oriented features of good writing concern the writing characteristics writers may employ when producing a written text. These of course are not often directly known by those assessing student writing.

Table 4-2 Supplementary features of 'good writing' mentioned by teachers

Category	Sub-category	Teachers' comments
Product-oriented	Content interest	T1 James: "being interesting is something in my opinion is fairly important".
	Lack of negative L1 transfer	T5 Gena: "negative transfer from students' L1. It is common between Chinese students and Arab. Sometimes I do understand what they want to say but sometimes I do not as it confuses me".
	Appropriate genre	T6 John: "the student should meet the requirements of the genre. My criteria depend on the task, if it is descriptive, if it is narrative and so on. My students know how to write in different genres".
		T2 Bob: "my emphasis depends on the writing genre. My scale differs according to the genre".
		T4 Sam: "presenting it [writing] in a different way, but may be with a different topic each time, but asking them [students] to write to a system and they're doing that".
	Presentation (handwriting etc.)	T5 Gena: "There are many features I think I consider them as very important for instance; lexis and vocabulary, collocation and also presentation in which I mean handwriting, layout ..."
	Context / Audience awareness	T3 Zain: "the student should know what the teacher expect them to write. I always tell my students to put themselves in the place of the reader, what do you want to read, what knowledge should you share?"
	Not necessarily Length	T2 Bob: "Achieving minimum words is not important as answering the question. If the ideas are clear and well organised and the student answer the question, this is enough for me".
Process-oriented	Planning	T4 Sam: "good writing should follow the given rubric and be planned according to it".
		T5 Gena: "I always tell the students: make sure you plan your essay, spend a few minutes making notes, just planning it in advance before you start writing it and make the necessary changes according to your plan. It takes some time but helps them to progress well in their writing".

Table 4-2 shows that teachers' perceptions of the important features of good writing not only covered writing as a product, but also seen as a process. Product related criteria were more commonly mentioned, however.

In terms of writing as a product, rhetorical features were stressed. Three teachers emphasised the importance of the ability to produce a text that corresponded with a particular genre, as each genre has its own writing conventions and fulfils different purposes. T6 John, T4 Sam and T2 Bob added this feature to the predefined features given to them. It was even commented that every genre has its own rating scale. Consequently, this feature has an impact on the quality of any written text and as a result it is considered to be important. As T4 Sam put it, the student has to write to a system, by which he meant knowing what language to use for a given type or genre of text. T4 Sam further explained that students need to know what kind of vocabulary to employ to serve a specific text.

> *If it is a descriptive text, the student needs to use descriptive language and include adjectives*

We will see later (4.9.3) how this is reflected in Sam's practices as a key feature of his observed rating style.

Another additional important feature of 'good writing' is interest of the content. What is interesting may differ from one reader to another, however, so arguably using this as a criterion for 'good writing' may lead to subjective assessment based on teachers'/raters' different conception of what is interesting. This comment corresponds with findings presented in previous studies that raters vary in their perceptions of rating criteria due to their different personal and professional backgrounds (e.g. Connor-Linton, 1995a; Lukmani, 1997; Cumming et al., 2002; and Eckes, 2008).

Another feature that may affect the quality of a written text is negative L1 transfer, according to T5 Gena. She commented that some students are affected by their L1 background negatively in their writing, which apparently affected the message that they want to deliver to the reader, as each language has different writing conventions and sometimes different structure. Although some teachers did not mention the issue of language transfer in the general interview session, we will see later that this issue appeared sometimes in their think aloud while they were actually evaluating their students' writing (4.6). Furthermore, L1 transfer was recognised as also being able to be a positive feature, in T1 James' eyes, if the L1 of the students has the same system.

> *Those kinds of mistakes do not normally happen with the European language speakers.*

Collectively, some teachers' comments suggest that if the students negatively transfer the way they express their message from their first language to a second language, this message might not be conveyed if the readers do not share knowledge of the first language. In this way, the students may fail to communicate in the second language. This point corresponded with previous studies which suggest that essays written by students from different L1 backgrounds differ in their linguistic, stylistic and rhetorical characteristics and raters consider this as a factor that affects ESL/EFL essay quality (Scarcella, 1984; Park, 1988; Reid, 1990; Frase, Faletti, Ginther & Grant, 1999; Hinkel, 2002). In contrast, if L1 transfer works well in the second language (positive transfer), transfer could be a characteristic of good writing.

Presentation was also considered one of the significant features that affected the quality of a written text. By this, T5 Gena meant handwriting and layout in her comments. She further emphasized this feature when she mentioned that:

> *If the handwriting is bad I am afraid I feel frustrated and cannot judge properly.*

In contrast, however, other teachers were more flexible in this matter. Some teachers even preferred the students' handwriting to typed text. In this respect T2 Bob added:

> *I prefer handwriting […] I find it much easier than something that's actually printed on computer. Maybe also people when they do it on computer don't leave enough space for correction […] and I think psychologically, if it's printed, it gives the impression that it's going to be right somehow; whereas if it's handwritten, it psychologically inclines you to think it needs more correction. But this is quite profound really. But I enjoy the handwriting more.*

T2 Bob therefore shares with T5 Gena the view that handwriting is a criterion that he considers, though he takes the opposite view of it to T5 Gena as far as its evaluation is concerned.

The issue of the audience, or what Hyland (2003) refers to as context, is another feature that was mentioned in the course of the interview. Only T3 Zain however highlighted this feature as an important matter for the student to pay attention to. In this regard, he emphasised that he informs his students constantly to write according to the expectation of the readers. Such a concern may be seen as part of a communicative approach to writing, since it emphasises that the purpose of writing is to convey information and opinions, not just to demonstrate language ability (cf. Zakaria and Mugaddam, 2013).

There is one more feature that concludes the list of extra product features, which is the length of the text. Regarding this issue, T2 Bob emphasised that he may accept the text as good if it is well organised and fulfils the required task and is of the required length, but he may be flexible if the length is the only problem with the text. In other words, he does not regard length as a key factor in 'good writing'.

In terms of writing as a process, the data show that teachers like T4 Sam and T5 Gena tended to identify 'good writing' as on-going activity that requires cognitive processing in order for students to plan their writing. They did not however mention other possible

process features such as that good writing was writing that writers had revised well, edited, written multiple drafts of etc., and had not written in L1 and then translated. Thus, they fell short of the idea that they considered "cognitive process as central to writing activity and in assessing the need to develop students' abilities to plan, define a rhetorical problem, and propose and evaluate solutions" (Hyland, 2003, p.10). There are many possible reasons for this, such as the difficulty for a rater to infer information about the writing process purely from the product, which is often all that he/she has to go on. Even though in our study the writers were students in the classes of the teachers rating them, the teachers would not have been able to observe and remember in detail the process that each writer followed when writing, even if they did it in class.

Overall, the comments provided by the teachers concerning the features of good writing demonstrate that they tend to value features of rhetorical structure and content more than lexicogrammatical features. Furthermore, two of the teachers at least clearly perceived good writing to be determined by a mixture of product and process elements. The findings suggest that the teachers, collectively at least, perceived 'good writing' as complex and multi-faceted and requiring a combination of various macro and micro skills to master.

The features of 'good writing' identified by the teachers are consistent with Hyland's (2003) five key features of writing knowledge:

> "Writers need to gain control of five areas of writing knowledge to create effective text: knowledge of the ideas and topics to be addressed (content), knowledge of the appropriate language forms to create the text (system), knowledge of drafting and revising (process), knowledge of communicative purposes and rhetorical structure (genre) and knowledge of readers' expectations and beliefs (context)" (Hyland, 2003, p. 113)

To sum up, the teacher between them showed awareness of all the key aspects of 'good writing' that are widely recognised, so might be expected to take these five features into consideration when actually rating compositions, which we examine further in later sections.

4.3 Rating scale and general criteria of the teachers, and their sources

We next turn to **RQ2**, which was also predominantly answered from the general interview:

> **RQ2: What rating scale and criteria do the teachers claim to typically use when rating student writing, and what are their sources?**

It was a vital issue to ask the participants about the scale they employ during their evaluation of their students' compositions, since they are not required by the IA to use a specific rating scale and criteria for assessing practice compositions, and, in order to replicate this situation, they were in our study also left to choose their own scale and judgmental criteria.

4.3.1 Scales and criteria in relation to their sources

First, regarding the scoring scale that the teachers adopt, they were asked about the types of scales they employed and how they used them and from where they derived them. The interview revealed that teachers in the same course employ different kinds of scales. One teacher (T 4 Sam) used a 33-point scoring scale, another teacher used a 10-point scale. Most of them did not spontaneously report evaluations with a number scale at all, even when in part using IELTS criteria, but simply a subjective and informal adjective scale employing evaluative words such as *good*, *bad*, *nice*, *interesting*, *poor*,

etc., especially for individual criteria. As T5 Gena said, referring to scale she used to summarise her rating (and report it to students):

IELTS is different, the scoring. I don't do IELTS marking

Table 4-3 Sources of criteria that the teachers employed

Source / Teacher	IELTS criteria in full	IELTS in part	CEFR	EELP Course system	Personal features
T1 James*	x				
T6 John					x
T5 Gena*		x		x	x
T3 Zain*		x			x
T2 Bob*		x	x		x
T4 Sam					x

* Taking IELTS class

As for the sources of the criteria used, it can be seen from Table 4-3 that the majority of the teachers (five teachers out of six) employed self-chosen criteria, and some also IELTS criteria. This finding confirms that the teachers in the study have a high degree of autonomy in deciding on the scale and criteria to be applied. In the interview, all teachers highlighted that the institution allowed them the freedom to choose their criteria and adjust them according to the students' abilities and the course requirements, which they perceived as being to meet students' future academic needs. The teachers also claimed that they were free to develop and choose their own writing tasks without obtaining permission from the course leader.

Table 4-3 indicates, also, that several teachers drew on criteria from a high stakes test, which were those of IELTS, since all the teachers except T4 Sam and T6 John were taking IELTS classes. In this case, the IELTS criteria are: task achievement, coherence and cohesion, lexical resource, grammar range and accuracy.

James' explanation for this choice was simple:

> *Because it is IELTS class, I am using IELTS criteria.*

While adopting IELTS criteria, and uniquely awarding a score on the IELTS scale at the end, James however did not implement the full formal IELTS scoring procedure of scoring each subscale and calculating the overall IELTS score from that.

The other three teachers taking IELTS classes did not adopt such a simple solution. T5 Gena for instance used her own mixture of criteria, largely based on IELTS criteria with more features added. She referred to 'common sense' and 'instinct' as her basis as well as general memory of IELTS criteria:

> *In fact, we had for this class ...we had a printed sheet with criteria from ELLP. But I did not follow it strictly, because I prefer to go with my own instinct. Yes, I look at it, but I think a lot of common sense. If you been doing this for some time and I am familiar with IELTS although I am not an examiner, I am familiar with what is required.*

The ELLP criteria referred to here are in fact IELTS criteria with one or two more criteria given to the teachers for information, by the leader of the ELLP course, without requiring the teachers on the course to use them. These criteria were: content, answering the question, cohesion, vocabulary, grammar, punctuation and paragraph structure if it is an essay. So, they were more or less IELTS criteria using different terminology.

T2 Bob also referred to IELTS but clearly did not use the IELTS criteria in their precise form, nor report scores on the IELTS scale. He used a scale of his own and referred to his personal impression of what a writer at B1 level would be 'expected' to produce. In effect, then, he is referring loosely to the CEFR scale and criteria (for B1 level) rather than IELTS ones:

> *My criteria in this case, because it's an IELTS B1 class and there's an irony there because IELTS is not for B1, IELTS is for B2 and C2. So how can we evaluate this in a way that's going to be positive for the students? Well, I try to assess it on what I expect someone at B1 to be able to produce. So, my scale is a six-point scale and the expected level is third from the top in that. And the fourth from the top is one which is just below and I would expect most people to get ticks in the "as expected" or "just below".*

In another comment, T2 Bob again shows that he is using other sources than IELTS:

> *It's more my own than from the centre, to be honest, but obviously, I am referring to knowledge that I have acquired.*

Reference to past knowledge and a general impression of a target level that the writers should reach is close what has been referred to as 'construct referencing' of assessment. However, in our study raters like T2 Bob are working each with their own personal construct, not one agreed among all the markers on the course, acting as a community of practice (cf. Wiliam, 1994).

The teachers' high level of use of personal criteria employed in their assessment may reflect the light level of support and control provided from the course program leaders/management. Indeed, this was confirmed by a comment from a course director who endorsed the teachers' claims about using their own criteria:

> *Each teacher has his own criteria and for that it may even be based on the overall class, the strongest students in the class, and the weakest student and the percentage comes from that. Other teachers have set criteria they look at. And they look at vocabulary use like in the IELTS for example […] Most teachers I guess they do not all do. It is not standard criteria we all use. We have our own criteria. In all honesty, some teachers might just give a percentage without criteria it depends on time but if there is time most teachers will check their criteria.*

While this agrees with what the teachers said with respect to using their own criteria mixed with elements of IELTS, the program leader here also mentions two practices

that our sample of teachers themselves did not report in the general interview data considered here. One is the idea that sometimes they base their scoring relative to the performance of the best and worst students in the class, effectively in a norm-referenced way: this is not the kind of way that scales like that of IELTS work, having as they do absolute, criterion referenced, definitions of scale points. The other is that the teachers may in fact not use separate, analytic, criteria at all but go straight to an overall score, in effect rating holistically. These themes are considered further in later sections.

4.3.2 Scales and criteria in relation to the pedagogical function of the assessment

Apart from deriving their scales and criteria eclectically from other assessment sources such as those above, we could interpret teacher choices also in another way.

Something that all the teachers seemed to share was that their scoring was in some sense analytic rather than holistic (chapter 3). For the teacher who closely followed IELTS practices this was inevitable, as that system requires separate ratings to be made of task achievement, vocabulary, etc. and an overall score to be derived from that by rule. However, even those teachers who awarded no numerical scores at all also, as we observed later in practice, seemed to be applying adjective scales to different aspects of the writing separately and then intuitively summing those up in an overall adjectival rating of the whole composition. Aside from the influence of other analytic scoring schemes such as IELTs or more local ones, which they had been exposed to, a further reason for this could be as follows.

Most teachers also gave detailed written feedback. Although the nature of this feedback is beyond the scope of our study, so what teachers wrote on scripts was not analysed, this is of course consistent with the assessment task that we asked them to do. We had

not asked them to assess the scripts as if for an examination, where often scripts would not be returned to students so there would be no point in writing feedback on them. Rather we set them to rate them as practice compositions done as part of normal class teaching, so, although we did not prompt them to supply detailed feedback, that would be common practice, especially in the form of feedback on errors at various levels, so as to help the students improve (typical of formative rather than summative assessment). Now if teachers are looking at a script at various levels in order to give useful feedback to students, with a teaching more than a testing purpose, then it follows naturally that their criteria for any rating or score may also involve similar aspects to those that they give feedback on, and that their overall assessment will be in some way be derived analytically from judgments of each of those criteria separately.

Thus, the use of an analytic scoring method may be due to the fact the teachers are concentrating on developing the students' writing regardless of their overall result. In the context of this research, where teachers would perceive that the students are keen to pursue their higher education by developing their writing, the teachers would naturally choose the assessment and feedback method that best suits this purpose. Indeed, to prompt their students' development in writing skill, the teacher may believe that giving detailed feedback is actually more appropriate than the teacher applying strict criteria to obtain an overall score.

In this extract, in fact Gena T5 overtly reflects the above ideas of evaluation when she speaks of her usual attitude in handling student composition papers:

> *I am here to help students to write better more than to put them in exam conditions where they expect a score.*

This was confirmed by what she said that she stated to them:

> *So, I always tell the students at the beginning of my writing course "I'm sorry I can't tell you what the score would be on IELTS. However, we are working with IELTS criteria. Just follow my comments"*

In other words, she has gone to the extreme of not supplying an overall score at all. For her, what matters is that the students see the information from the analysis itself that would precede awarding a score in normal analytic rating.

More generally we could say that while holistic rating/scoring is seen as awarding an overall rating without prior analysis, what we see here is a form of analytic rating/scoring which consists of doing the analysis but not in fact performing the process of deriving from it an overall score at all, or at least not one that is given to the students. This is not a variety of analytic scoring usually discussed in the assessment literature.

Ratings, whether as scores or evaluative adjectives like *good*, given for distinct categories such as organisation, task completion, cohesion & coherence and mechanics can not only help students, especially if accompanied by further feedback, but also help teachers to better determine the students' writing proficiency levels diagnostically, i.e. with an understanding of precisely where the strengths and weaknesses lie in a particular student's writing. This in turn can guide the teacher's class instruction to improve the students' performance (A. D. Cohen, 1994; Hamp-Lyons, 1995; Bachman & Palmer, 1996; Shaw & Weir, 2007). By contrast, holistic rating cannot contribute to measuring students' strengths and weaknesses, especially for those who obtain mid-range scores (Hyland, 2003). Thus, it is perhaps also for their ability to measure students' performance in detail that analytic rating criteria may be preferred by teachers.

The above account shows that our teachers for the most part perceived their role as being teachers first and raters second, and chose analytic criteria accordingly. They also

did on occasion, however, refer to a third possible role, that of ordinary lay reader, and the tension which they experienced in separating that from the teacher role. This tension showed itself in issues of choice of criteria and of balancing their personal reaction to students' scripts and a set of criteria. This was not discussed in the general interview but a good example that illustrates this tension is in T5 Gena's data responding to a particular essay:

> *I mean, obviously, I'm reading this as a teacher: I'm marking it, evaluating it. But somebody who's just reading it, if it were an article in a newspaper or something, you'd get a little bit fed up, you know? […] I mean, if I was reading this, as I say, online – an article, a newspaper or something, online newspaper – it starts to get a little bit tiring when you see mistakes all the time. And you lose – it detracts from what you're reading; you can't focus on the message because the mistakes get in the way. But I'm trying to see beyond that, because obviously, I'm the teacher; I'm not just any reader. But yes, she needs to... She's got good control of the language, I can see that, but what she needs to do is sort of fine tuning. She needs to perfect it, really, and these are the details. But she will get there, because I know her and she's a strong student. So, I'm not worried about her at all.*

Interestingly, this teacher seems to be faced with a conflict between how to respond (i.e. what criteria to use) as a general reader versus as a teacher/rater (roles that she does not separate in this case). In the end, she opts for the perspective as a teacher. This attitude was found to be common among raters in Cumming's et al., study (2002). Even experienced teachers, who are prepared with rating criteria that serve their course requirements, may mediate judgment on written text using their perception and personal interpretation (Lumley, 2005). This theme also links to the criterion of audience awareness which we identified in 4.2, and evidences that some of the teachers are interpreting writing in a communicative rather than just a language/skill based way.

4.3.3 Variability within teachers in criteria used and their weighting

Consistent with the previous section, the data further showed that an individual teacher's evaluation criteria, whatever their source, did not necessarily remain the same across scripts. There were relatively few statements, either in the general interview or when talking about the specific essay evaluations, where teachers claimed to have some general policies about criteria that applied universally when they rated compositions. A few examples of such choices set in stone come from T2 Bob:

> *First I always look if the student answer the question, and the organization. Grammatical range is less important for me than grammatical accuracy and organization. I think if I have those I think then I can develop the others.*

Here T2 Bob claims two criteria that he always considers, regardless of essay topic, student etc., and he further declares a universal weighting applied within grammar to value accuracy above range (and his practice largely confirms this, 4.9.1).

Far more often, however, we found statements implying that teachers saw themselves as operating with criteria used contingently, not universally. That is to say that the criteria used at all, and the weightings of those that were used, relative to each other, depended on factors unique to particular essays. Those factors mostly came down either to the writing task/prompt involved for that essay, indicating the genre, or the nature of the writers. T3 Zain for instance refers to a criterion of each of these general types when he says:

> *We have to be careful in choosing suitable criteria because each genre has its own criteria, although there are some common features among different genres. So, we need different criteria with different approaches. Students' level is another indicator I would say.*

In this view, it is clear that T3 Zain thinks essays of different genres need their own criteria to some extent, and so do classes of students of different proficiency levels. His practice (4.9.5) does largely support this.

T2 Bob also stated that his scale differed by genre:

> *My emphasis depends on the writing genre. My scale differs according to the genre.*

Again, T4 Sam said:

> *In this particular essay, I am keen on grammatical range and lexis.*

Here T4 Sam is again drawing attention to weighting of his criteria being dependent on the writing task type (topic, genre etc.). Elsewhere, however, he interestingly admitted that, despite such contingencies, vocabulary tended to emerge as an important criterion regardless of genre:

> *But, depending on what type of paragraph they're writing: this is a descriptive paragraph, they need to use adjectives; if it was a compare-and-contrast, then they'd have to use the vocabulary that matches that paragraph; if they were writing opinion, then there would be an amount of vocabulary they need to write about their opinions. So, I would say, if I'm honest, the one where they can score the most points in the hierarchy is vocabulary.*

In the actual rating reported in the present study there was little chance for teachers to evidence shifting scales and criteria due to difference of genre, since each teacher was typically rating essays all written to a prompt eliciting the same genre (though that genre differed between teachers, either descriptive or argumentative). We will see later however that in practice the teachers did sometimes value features of good writing differently based on the individual characteristics of each writer.

These comments show that the teachers weighted writing features differently on different occasions and might focus on one or two criteria outlined in their rating guide or use their own internalised criteria. These findings again imply that the teachers in this context are free not only to make their own decisions on what criteria they use in their course, but also how to vary criteria and their weighting from essay to essay.

It is apparent however that some teachers experienced some tension between being objective and subjective in choosing suitable criteria and in using them consistently or not. As T6 John, for example, said:

> *Fixed criteria can be more reliable than our own ones and sometimes it is fairer to use them. But knowing the students' needs and level, using a fixed scale can lower their grades from the overall grades they deserve. I think personally, if we have set criteria, it can tie our hands. Sometimes, I prefer a mixture of criteria, where I can be present. In this class, we are not testing students' proficiency but we are trying to develop their writing skill in a more pleasant way, and gradually prepare them for IELTS.*

This comment is of great interest in relation to teachers' use of criteria. Although, the teacher emphasised that using a fixed rating scale and criteria can maximize the reliability of the assessment (highly valued by testers, as described in chapter 3), on the other hand, he in effect conceived it as lowering the validity of the assessment, though he did not use that word. For him, as for others we saw earlier, the real purpose of the assessment is student improvement not obtaining an accurate measure of their ability, and from this comes his preference for variable criteria and the right to choose the appropriate criteria that accommodate his experience and knowledge of a particular context not just universal criteria. In terms of the types of validity which we reviewed (chapter 3), he could be seen as favouring consequential validity over the other types of validity which are more concerned with true measurement of a target construct of writing proficiency.

Another comment by T4 Sam gave different arguments in favour of variability:

> *What I think is, the truth is it's not an exact science. You can't... If you give yourself exact criteria to work to, it's not really that straightforward, really. If you're marking for grammar, then fine, mark everything for grammar; but then you're not looking at the content, you're not looking at the structure, you're not looking at organisation. And I think with writing there is this level of subjectivity which means it's a little bit personal as well. And when you've worked with students, because this is not a credit-bearing module, I can make my own criteria and I can be a bit more generous – a bit more generous. And I think by being a little bit more generous, it can motivate the students.*

Here T4 Sam argues first that the strict fixed approach requires the teacher to focus rather artificially on different criteria separately, implying that this is not very natural or easy to do, and so perhaps not in the end as accurate as it might seem (contrary to T6 John's claim of *reliability*): perhaps the existence of subjectivity proves to be inevitable.

He then makes the second point that by being selective in his criteria and their weighting, on a subjective basis, he can end up giving students a better score, which has the benefit of encouraging them. In this respect, he pays attention to the affective side of the writer, perhaps also see in the use of the word *pleasant* by T6 John in the previous quote above. Positive motivation and reduction of writing apprehension are both seen in the literature as affective aspects of the writer that need to be enhanced alongside more cognitive and metacognitive skills.

In both of these arguments T4 Sam is then again making a case that, in the context of the current writing course, assessment does not need to be rigorously fixed and analytical in pursuit of high levels of accuracy, but rather more variable, naturalistic and focused on student progress, in this case by working on the affective side of the student.

Teacher views such as those above suggest that our teachers are aware of the process of their synthesis of personal judgment and experience in their assessment, and of the context-related complexity of employing suitable criteria. They see that merely having a set of criteria or a fixed rating scale might not guarantee the true fairness of the assessment. Teachers' own decision making seems to have the priority in evaluating writing in their courses at any rate.

4.4 Teachers' training in writing assessment and views on such training

In this section, we answer **RQ3**, again predominantly from the general interviews.

> **RQ3: What training have the teachers received that is relevant to assessing writing and what are their views on such training?**

Writing assessment training is not only clearly crucial for writing testers/examiners (ref to your lit), but also a vital part of EFL/ESL teachers' general training for their role in classroom teaching and assessment (chapter 3). Teaching and assessing writing are not an easy task for teachers, as we saw in the previous sections, so training may help them to gain principles and practices that are needed in order to guarantee the suitability of assessment. In the general interview the teachers were therefore asked if they had had any such training course, and, if so, their experience of it.

It is worth mentioning that discussing the issue of training is not at the heart of this study. However, teacher experiences and attitudes with respect to training might help to provide explanations behind a particular rating behaviour or decision making process for some teachers. Hence this section may help inform later sections.

4.4.1 Relevant training received

Five teachers reported that they had either had a comprehensive writing assessment course or a general course that included coverage of assessment, or both (Table 4-4). Only one teacher stated that he never had any kind of training in this area. The type of training that each teacher had received varied enormously from highly professional specialized courses such as IELTS examiner training, and courses on how to teach and assess writing for exams such as CAE and FCE provided by Cambridge University, to quite local and less formal training in the form of departmental or institutional tutoring on assessment. The length of training varied between one day and five days.

Table 4-4 Training courses that teachers received

Name	Courses/Training courses
T1 James	IELTS examiner training courses
T2 Bob	Cambridge main suite examinations: marking of Cambridge Advanced English (C1 level) and First Certificate English (B2 level)
T3 Zain	General training in language assessment.
T4 Sam	CELTA
T5 Gena	DELTA
T6 John	PGCE and TESOL.

The emphasis on different components of training was different in different training courses. T1 James gave full report about the IELTS training course:

> *In terms of formal training – I'm referring again to my IELTS training. So, when I first trained as an IELTS examiner in 2007 we had a four-day course. Two days was for writing and two days was for speaking. So we had two full days of practice in marking writing. So that would begin with being shown samples of student writing and being told "OK, this is a 7 for task achievement, a 6 for coherence and cohesion, a 7 for writing and a 7 for grammar", and looking at it and analysing why it rated at a 7 or a 6 or whatever. And then slowly, slowly we would look at all the different criteria at each level, according to the IELTS system, and sort of fully understand them, look at examples of writing and sort of explain why this was a 6, why this was a 5.*

> *And gradually we were allowed to actually practise marking scripts and to check whether our scores were the same as the official scores that the experienced examiners had agreed upon. And then basically what happens is, every two years for writing, we are re-certified, so we have to re-train. We have four hours of training in the morning, and then in the afternoon we have to mark a load of scripts. And again, our scores that we give are checked against the kind of official scores that were agreed on by the top IELTS examiners in the world. So, every two years we're kind of re-examined to check that our marking is at the right level.*

At the other extreme, one of the expert teachers (T2 Bob) highlighted that the kind of training he had was not even particularly on assessment, let alone writing assessment, but rather something general:

> *I have been on some courses which included an element of assessment training. It's not been specifically an assessment training.*

T6 John gave more details of such a background:

> *Yes. I mean, um, well I suppose from my qualifications, my teaching qualifications are PGCE, which is postgraduate certificate in education, where they give us very basic training, really, because that's designed – PGCE is not designed specifically to teach writing. It's basic because, of course, with any training you're asked to look at grammar and vocabulary. I also have the TESOL, which is a level five certificate in teaching English to speakers of other languages, so that's the specific, it's a one-year course in teaching English to – it's postgraduate – teaching English to speakers of other languages. So that focuses much more on marking writing – things to look for. And of course, a lot of things tend to fall into patterns. I mean, they tell us quite obviously, right from the start, a lot of Arabic writers will tend to spell a lot of words the other way around, because of the way Arabic readers read. So, the training does incorporate some nationality issues as well.*

Clearly such training had none of the rigorous familiarisation and practice with one assessment system that is the key feature of IELTS training. On the other hand, it incorporated attention to errors typical of learners with different backgrounds, which is

an aspect of writing assessment much more suited to a teacher doing informal assessment of practice assignments, for formative purposes, accompanied by full feedback.

In general, the data showed the majority of teachers are not well trained in professional writing assessment up to level at which they teach. Most of them either have had general language assessment training or not been trained at all, except for one teacher who trained to be an IELTS examiner, which arguably is over-trained for the sort of teaching rather than testing oriented assessment on the course they were teaching. The training within the institutions where they were teaching does not take the form of courses but more of professional development in the form of occasions where the teachers meet and discuss a topic such as 'writing assessment'. There seemed to be a lack of dedicated professional training support specifically for writing assessment offered to the teachers inside the institutions.

4.4.2 Attitudes to training

Overall, there were not lengthy comments about the training, especially the local training. However, the teachers do provide more favourable than negative comments on training. Some teachers' favourable comments are summarised in Table 4-5.

Table 4-5 Teacher positive comments on the usefulness of training

Aspect of teacher that benefited	Sub-category	Excerpts
Rater feelings (affect)	Effect on confidence	T1 James "I can feel that training boosts my confidence in assessing my students' compositions."
	Basis for the effect	T3 Sam "You feel that your assessment is based on a solid basis."
		T6 John: "Training gave me the chance to reflect on my experience in writing assessment in a more positive way."
Increased rater knowledge of writing assessment (competence)	General	T2 Bob "Despite the little training I had, it was useful in giving new ideas and new information about marking skills.
	Criteria	T5 Gena "Before training my judgment was sometimes based on two or three features."
		T3 Zain "My focus during writing correction as well as my understanding of some aspects of writing features, have developed after training."
	Rating scale and criteria	T1 James: "This class is an IELTS class, training clarified to me how to use the rating scale and how to apply rating criteria."
	Rating process	T2 Bob "Training in all aspects of teaching gives appropriate procedures to follow"

Table 4-5 shows that many of the teachers believe that training can benefit them affectively/emotionally, as well as by improving their knowledge so as to better guarantee their assessment appropriacy.

Teachers mentioned a variety of effects on their assessment skill. This included not only better understanding of scales and criteria, e.g. widening their attention from two or three features during their rating, but also the procedures of doing the rating, showing awareness of it as a process. As T1 James pointed out, frequent training (every two years in his case) also helped the trainee to keep up to the right level and would focus on the need for inter-rater agreement, and help improve that (Weigle, 1994, p. 200). He clearly had in mind, however, professional exam assessment (e.g. IELTS) of the type he described in 4.4.1 rather than everyday informal classroom rating of writing. By

contrast, T5 Gena, referring to departmental training, essentially viewed training as useless due to it not covering anything new:

> *Sometimes, at the departmental level, training is something repetitive, I am not bothered if not attending it.*

T5 Gena refers here to the sort of regular meeting described at the end of the previous section.

The different attitude to repetitiveness of training of James and Gena we interpret as related to a difference between how training is perceived in relation to high stakes professional exam marking, compared with everyday classroom evaluation of the type in our study. The former is seen as needing to be repeated every few years, however boring it may be, so that examiner ratings are as accurate and reliable as possible. Such repetition is not seen as essential, however, for the less sophisticated internal low stakes departmental marking of class compositions undertaken by teachers like ours, however, perhaps due to its different, non-testing, purpose.

It is also noticeable that the beneficial effect of training on rater feelings in the form of confidence was referred to, not just the effect on rating ability. Teachers were aware of the importance of this dimension and not just that the teacher needed to have good rating skills. In particular, one teacher, John, welcomed the positive frame of mind that came from reflecting on experience in a training course.

This last point leads to a further interesting theme in the responses, which was the connection between experience and training. John, as just stated, seemed to welcome the opportunity to merge both experience and new knowledge gained from training in assessment practices. T3 Zain, however, commenting on the usefulness of training for

novice teachers, incidentally demonstrated that he believed that experience can substitute for training:

> *I am not an expert in teaching or writing assessment therefore, I reckon that training could be useful for new writing teachers. It would help them to build their assessment knowledge that is not yet acquired by experience.*

Indeed, for this reason some teachers saw training as something less important. T2 Bob voiced this more bluntly:

> *With the experience I have, there is no need for training. I also go with my instinct of experience. It depends on my students really, whether it helps me with my own students or not. For someone like me […] training definitely is not for me, I believe.*

In this comment, the teacher is clearly referring to training for the sort of everyday informal classroom rating of practice compositions that our study was focused on, not professionally rated IELTS exams. Therefore, T2 Bob believed that training might be relevant only if it corresponded with the particular type and level of students that the specific teacher taught (compare 4.3.2). Otherwise he claims that his long experience of that context gives him all the intuitive expertise he needs. T2 Bob voiced this in more detail as follows where he seems underestimate the value of training compared to his prolonged experience by not mentioning it.

> *I have been working on this for 19 years, so […] I think that once you've seen a few hundred things, then in some ways you're able to recognise certain patterns, and therefore you are able to assess things pretty quickly. For example, in the main suite examinations, like first certificate, to give a general assessment I can usually do that […] I mean, certainly to look at it in detail and refine it would take more time, but I think, as I say, when you see hundreds of these you very quickly get a picture of them. The IELTS scripts do take a bit more looking at.*

It is interesting however that he seems to identify assessment expertise with how quickly ratings are done rather than how accurately, or how usefully for teaching purposes.

It could be argued that James and Bob represent two extremes of beliefs about teacher professional development, with respect to rating of writing (cf. Wallace, 1991). T1 James has in mind rater development as promoted through top down training by experts in a universally approved research-based rating scheme. This is what Wallace calls the applied science model of teacher education, and of course for an exam like IELTS it is the only realistic rater development choice. T2 Bob by contrast sees rater development as occurring bottom up, from teacher experience in a particular context where the teacher teaches, not from experts, and as leading to a rating scheme that might be unique to that situation, tailored precisely to local needs and students. This is close to what Wallace (ibid) calls the reflective model of teacher education, since, although T3 Zain and Bob do not quite say it, experience does not teach a rater anything unless they reflect on it and learn from that.

T6 John however, as we might guess from his earlier comment, saw training and experience on more equal terms. Having described useful features of his training he added:

> *But of course, the other thing is, from an ongoing point of view, I mark every week so it's an ongoing learning process.*

Thus, he saw an interplay between training and learning from experience occurring alternately and benefiting from each other.

In conclusion, it appears that the flexibility given to the teachers in designing rating criteria for the course they were teaching also allowed them to be more independent of

training and rely on their own experience. Indeed, they would not find a training course for their own precise chosen scale, criteria and procedure precisely because they selected and designed those themselves. Moreover, the data showed that the teachers (perhaps with the exception of T1 James) were not concerned with agreeing with others (cf. the emphasis on reliability in testing). This outcome corresponds with an earlier study by Weigle (1994) in which the flexibility that had been given to her informants allowed them "to be more independent in their rating than if the criteria for forcing a third reading of an essay were more stringent" (p.214).

4.5 Teachers' actual rating criteria reported used in their assessment of essay qualities

We now move to answer **RQ4** concerning the criteria teachers were found actually using when rating specific scripts, based on the TA data and the immediate follow up interviews.

RQ4: What are the most important qualities that the teachers look at in practice when they are rating their students' samples? (criteria used and their weighting)

What they have to say about this of course is prompted primarily from their reading and interpretation of the essay texts, the process of which will be looked at in more detail in 4.8 and 4.9. In particular, our coding scheme made a fundamental distinction between evaluation of features in the essay text, using criteria (=judgment), and what the teacher did just to understand the text (=interpretation). The latter was often a prerequisite for the former so occurred first: see 4.8.2.2 for examples.

It must also be noted here that while many criteria were named in standard and transparent ways by participants, using labels such as argument, examples, spelling,

collocation, linking word, word choice, etc., some criteria were idiosyncratic, as might be expected when teachers were, as we have seen, often not simply adopting the criteria of a standard scoring system like IELTS. One such was the criterion of consistency, mentioned a number of times by T4 Sam in the TA but not in his general interview. While this label could in principle apply to many things, careful analysis of the TA protocols showed that he used it specifically for repeated use of the present simple, which he focused on as a signal that writers were suiting their language to the genre of the essay, which was descriptive. This was clearest in this excerpt:

> *Let's have a look for his consistency and his tenses. "Was born", "is born". OK, so this is all in the present. "He is a good friend and we plan to be very good friends and take care of each other because we are" – yes, OK, he's actually remained very consistent throughout.*

4.5.1 Frequencies of mention of criteria

In Table 4-6 we list all the relevant criteria as we coded them, in the major categories of evaluations of language or of rhetorical and ideational features.

Table 4-6 Frequencies of references to different criteria in the TA

Codes	James	Bob	Zain	Sam	Gena	John	Total
Rhetorical and ideational focus Judgment strategies							
Task fulfilment of requirement							
Relevance	3	5	1	6	6	3	24
Task completion	4	2	2	4	2	1	15
Relevance, quality, appropriacy of argument	1	4	7	5	13	7	37
Identify redundancies	0	0	1	1	0	2	4
Organization							
Paragraphing	12	7	4	3	2	3	31
Overall organization	0	1	3	6	4	3	17
Main body	4	2	4	0	2	0	12
Introduction	9	5	11	1	11	2	39
Conclusion	5	1	5	5	7	1	24
Cohesion & coherence	11	8	1	6	9	0	35
Reasoning, logic or topic or text development	6	5	8	15	5	1	40
Total	55	40	47	52	61	23	278
Language Focus Judgment strategies							
Codes	James	Bob	Zain	Sam	Gena	John	Total
Vocabulary							
General comments	1	4	1	1	0	0	7
Errors	12	7	6	3	2	0	30
Choice, range, appropriacy	12	4	6	13	9	5	49
Rate language overall	9	4	6	8	7	3	37
Proficiency of language use	8	6	3	1	5	1	24
Mechanics							
Spelling	16	11	6	0	5	8	46
Punctuation	9	9	11	0	0	1	30
Capitalization	3	3	3	0	0	3	12
Syntax or morphology	5	2	3	3	3	1	17
Language Error gravity or severity	8	7	1	2	2	1	21
Language Error frequency	6	6	6	1	8	4	31
Clarity, confusion, comprehensibility of language	13	4	19	2	4	2	44
Quantity of written production	6	1	2	9	4	2	24
Total	108	68	73	43	49	31	372

It is immediately obvious that there are considerable differences between individual teachers, with James at one extreme making more than three times as many references to different criteria than John at the other. T1 James had had the most rigorous training, as an IELTS examiner, and this perhaps not only trained him to be more thorough when scoring essays in an IELTS class, but also gave him practice in talking aloud about criteria and how they were guiding him to a score which the others lacked. John by contrast we would regard as a competent rather than expert rater, and was not teaching an IELTS class but was not the teacher with the least training. His result may however be explained by the fact that John's writing task was different from that of the other teachers. Unlike the other teachers, John chose a simple picture description task which was suited to the lower level of his class.

In between the extremes, T2 Bob and T3 Zain are higher than T4 Sam and T5 Gena in references to language criteria, but the reverse is true of rhetorical and ideational criteria. T4 Sam for example makes no use of criteria related to mechanics. The reason behind this attitude was that he felt that the nature of the task, 'describing a friend', meant that vocabulary choice was more important than mechanics (see discussion in 4.3.3).

T5 Gena, on the other hand was the highest in the use of rhetorical and ideational criteria and lowest in use of language criteria. She had set an argumentative essay and she demonstrated that her interest was particularly in the content by saying for example, when asked about what made the essay good:

> *He answers the question – he misses something, as I said, but his argumentation is good.*

T3 Zain, by contrast was much more focused on language, especially punctuation and language clarity. This could be due to his background as a learner himself in Syria where teachers possibly focus on language use in writing assessment.

Looking at the specific code categories, we find that overall there are more references to language criteria than rhetorical and ideational ones unlike Cumming et al., (2002) whose informants were keen more on rhetorical and ideational focus than language focus. This difference could be due to the fact that in Cumming's study the raters were expert examiners marking IELTS essays not practice classroom writing. In our study only T4 Sam and T5 Gena showed this trend in favour of rhetorical and ideational criteria.

Of course, it must be recalled that the result from the teachers' self-report protocols must be seen as tentative. As we noted in chapter three, Nisbett and Wilson (1977) warn that a verbal report is not complete record of a person's thought processes. In particular, the absence of any particular phenomenon in a protocol is not evidence of its absence in actuality, since it is impossible to report on all of one's thoughts at any given moment. However, the protocols are revealing in showing some interesting differences between these teachers' rating criteria insofar as they were able to voice them (Weigle, 1994, p.207-208).

Looking at codes individually in order of frequency (Table 4-7), we can see that among the language criteria vocabulary choice and spelling were most commonly referred to, followed by overall ratings of language, then followed by three criteria from rhetorical and ideational focus.

The rhetorical and ideational criteria were best represented by criteria related to the reasoning and appropriacy of argument of the content, perhaps due to the fact that five

of the teachers used argumentative writing tasks. Of all the parts of the text structure, the introduction seemed to receive most attention. The reason could be that it was always the first part to be read, so perhaps received more mention simply because the rater was fresh to the essay at this point. Furthermore, it also could be due to the teachers treating it as an opportunity to build their initial impression which would ultimately colour their whole rating of the essay.

These findings may be compared with what they said in 4.2 about the characteristics of good writing in general. The two sets of findings agree with respect to the high result for discourse/rhetorical and content/ideational features, and the high placing of vocabulary, on the language side. However, the high position of spelling in the present section shows perhaps a gap between teacher beliefs, as represented in 4.2, where spelling came lowest, and their practice, as represented by frequencies of mention in the current section, where it comes very high. The result also matches Cumming et al. (2002) except with respect to the high placement of language features.

Table 4-7 Criteria in order of overall frequency of mention

Criterion	Frequency
Vocabulary Choice, range, appropriacy	49
Spelling	46
Clarity, confusion, comprehensibility of language	44
Reasoning, logic or topic or text development	40
Introduction	39
Relevance, quality, appropriacy of argument	37
Rate language overall	37
Cohesion & coherence	35
Paragraphing	31
Language error frequency	31

Vocabulary Errors	30
Punctuation	30
Relevance	24
Conclusion	24
Proficiency of language use	24
Quantity of written production	24
Language error gravity or severity	21
Overall organization	17
Syntax or morphology	17
Task completion	15
Main body	12
Capitalization	12
General comments	7
Identify redundancies	4

We noted in 4.3 that teachers did not necessarily use the same criteria for every essay. My analysis of comments made by teachers on the essays with respect to their criteria shows however that in fact in most cases the individual teachers maintain the same criteria in their detailed comments throughout their assessment. Moreover, generally, different teachers share more or less the same set of common criteria. As we see from table 4-6, out of 24 categories of criteria, two thirds were used by all the teachers. This is remarkable given that the teachers were not following an imposed set of criteria but were left to their own choices.

4.5.2 Wording of evaluation of essays using the criteria

Apart from which criteria were most and least mentioned, it is also informative to look separately at the positive and negative references to criteria. These comments include evaluative ratings that influence the marks or overall rating awarded.

In Table 4-8 example negative and positive comments are presented, summarized into four main characteristics that appeared to determine the teachers' judgments of the quality of their students' written essays. It is notable that organization and paragraphing were the features most mentioned that were associated with low scores. This result is not similar to Russikoff (1995) who found that when raters assessed ESL compositions holistically they paid attention only to "language use", which was seen as an ESL student weakness. It is perhaps closer to McDaniel (1985) who found that, in analytical rating, raters rated essays written by ESL and NE students differently in terms of the following three specific criteria: "content development and organization," "sentences" and "words".

Table 4-8 General characteristics that determined teachers' judgments of the quality of the written compositions

Characteristics	Positive comments	Negative comments
Task fulfilment	Task completed	Off topic and not complete
Organization & paragraphing	Well organised/ good use of paragraphs	There is an organisation issue and poor paragraphing
Language	Good language/ pretty nice vocabulary	Not advanced vocab
Cohesion & coherence	Very coherent and ideas are well linked together	No linking words

While it was not an aim of this study to look in detail at the numbers of positive and negative ratings associated with different criteria, teachers or scripts, we were wished to illuminate the ways in which teachers expressed their positive or negative ratings, in the absence of using scales of letters or numbers universally either to rate specific criteria or essays overall (4.3). Table 4-9 shows examples of how teachers worded their positive and negative ratings.

One noticeable feature of the way in which teachers talk verbally about their ratings in practice is that they often do not make clear the referencing of their evaluation. Examples of unambiguous references in Table 4-7 are rare. For example, where a teacher says *in terms of IELTS criteria he has done well* or *Not advanced vocabulary*, or *basic grammar mistake* we can reasonably infer that reference is being made to some absolute standards or proficiency levels (criterion referenced). If a teacher says *better than average* it sounds more like the script is being rated in relation to other scripts of that person or of the class (so norm referenced), not some absolute standard (cf. 4.3.1).

However, the vast majority of wordings are ambiguous: e.g. *the sentences are generally correct grammatically* does not tell us definitively what level of grammatical proficiency they put the writer at, since the sentences might be grammatically very simple or very complex; *poor spelling* could mean poor relative to the level of the group (e.g. poor for the teacher's class) or poor in some absolute terms (i.e. beginner level, A1); a bare comment like *Organization* is presumably negative, but again we cannot tell what referencing the teacher has in mind. As suggested in 4.3.1, the references may in fact often be to the teacher's personal image or 'construct' of what a student of the sort being rated 'should' be like at their level, so in fact more construct referenced. We shall see in 4.6 however that sometimes wider references accompanying the basic rating statements make this clearer.

Table 4-9 Summary of teachers' good and bad ratings in some students' compositions

Teacher	Sample	Positive	Negative	Other comment
1 James	1	-Advanced grammar. -Nice vocabulary. -Answers the question correctly. -Organization good. -In terms of IELTS criteria he has done well	-Chaotic sentences at the beginning -He needs linking words -Missing articles -A couple of spelling mistakes	Bear in mind he is only a B1 level[6]. He has not made as many mistakes as many of his classmates might have made. This is better than average.
	2	-Answers the question clearly - Supports her ideas -Sentences are complex, using subordinate clauses	-Poor spelling -There is an issue in paragraphing -Careless about her uncountable nouns -Vocab is not very ambitious. -Short intro and no conclusion	-Bear in mind their nationality. One of Grazia's strength was sentence complexity because she is Italian. So, when marking I expect that.
	3	-Answers the question completely. -Good paragraphing -Flashes of good language -Vocab and grammar good.	-Some confused bits -A couple of spelling mistakes.	-Really impressed by Ali as he recently joined the class, he is an intermediate student not high.
2 Bob	1	-The range of accuracy of vocab is pretty good. -The sentences are generally correct grammatically.	-Organization -Not approaching the question in the best way.	-Better than I might have expected from him to be honest, Amjad.
	2	-Paragraphing, good introduction, main body and nice conclusion. -Well organised. -Over-ambitious vocab	-Basic grammar mistake -Lack of care of building blocks.	-They know I am not happy if you have a sentence without a verb. This really irritates me.
	3	-Some good vocabulary -Large degree of acceptable organisation	-Completely off topic -Putting new subjects and new verbs after a comma rather than after a full stop.	
3 Zain	**1**	-Introduction and conclusion is ok -The idea in second paragraph is ok	-Grammar is horrible -Got the ideas but has not got the language to express them. -Frequent errors of passive. -Using *so that* at the beginning of every sentence.	-I keep saying this student should not be in C1B which is the advanced level. I would put him in intermediate or upper intermediate.
	2	-Good argument -Good grammar -The ideas are well presented	-She missed one paragraph. -Missed one important part of argument. -Under length	-Because it is grammatically good it does not matter if you write less than 250 words. -Arwa is exceptional.

[6] B1 level is intermediate level.

	3	-Good introduction -Grammar is also good	-Handwriting is not ok. He will suffer in IELTS if he writes with that bad handwriting -There is no thesis statement -Did not mention the second opinion of the task	-I am surprised because one of the teachers mentioned that he is bad in listening and speaking
4 Sam	1	-He keeps the tense consistent -He managed to support. -He is controlling the idea with the topic sentence. -His coherence is ok. -He used technical structure.	-He has not managed to summarise or conclude the composition. -Some spelling mistakes. -Some missing articles.	-I am going to revisit this to see if it is good as the other compositions or better.....when I read all the scripts. -He actually took on board what we have been teaching over the 5wks. -This is actually the best writing this student produced for me.
	2	-He used good topic sentence to start with. -Has got good control of the ideas. -His composition is so coherent.	-He messed up the second paragraph by using a rhetorical question. It does not really need it. -He wrote too much… he wrote more than 150 words and he has gone off the topic a little bit.	-However, I won't take a mark away. I would point this out. That is in IELTS test, this is where they would criticise you not being clear.
	3	-Stuck to the task -Stuck to the number of words.	-A bit too simple -No detail. -Simple structure	-I have decided to give her extra couple of marks because she stuck to the word count. I will take away marks from the other two scripts because they go over limit.
5 Gena	1	-Answers the question -His argument is good -Gives good sort of example -It is an interesting essay -Followed a good structure of IELTS task -Uses good language.	-One part of the argument is missing. -Incomplete task.	-He just felt he'd written enough or maybe he forget the question. That is what the negative is for me. Other than it is excellent essay. -He's making mistakes with countable nouns which is a common mistake and I keep telling them, this student in particular.
	2	-He has tried to be on the track -Language is ok -Uses good examples -Uses cohesive words and linkers	-A little bit weak -Misunderstood the question -His conclusion is very poor.	
	3	-A good degree of accuracy -Good grammar -Interesting	-Some mistakes like articles	

6 John	1	-The structure is good -The introduction, the conclusion and main body is pretty good -She picked up most of the issues that should be addressed in this kind of essay.	-There is a grammar issue. -Missing articles -Wrong prepositions -Spelling -Vocabulary use	-We have to bear in mind this is B1A, which it is not bad it is the same sort of problems I'd expect to find in the intermediate level so it is pretty nice piece of writing.
	2	-Good introduction -Uses nice vocabulary -Content is good	-Few spelling errors -Few plurals as well -Main body is really short and does not explain in as much detail as needed	

4.5.3 Summarising or combining ratings/scores for separate criteria into an overall score/rating

Aside from evaluating essays with scores or ratings by referring to separate features of the text, following chosen individual criteria, a key feature of any analytic scoring is the way in which the ratings of those separate features get to be combined into an overall score or rating. This is especially the point where differential weighting of criteria may appear (such as is built into some professionally made writing assessment systems, such as the Jacob's scale, which systematically weights mechanics much lower than content, for example). This is the core of what is often called 'decision making behaviour' in talk about assessment. Some of our data coded as Self-monitoring judgment strategies relate to this. Specifically, many of these examples are coded under code a. Summarise, distinguish or tally judgment collectively or code b. Main cause for given score. (for list of codes see 3.9.11)

We can see some examples of reference to this in Table 4-9. For instance, when T3 Zain says:

> *Because it is grammatically good it does not matter if you write less than 250 words.*

This is effectively indicating that, when arriving at an overall score or non-numerical rating, he puts more weight on the criterion of grammar than on that of word length.

Again, T4 Sam says:

> *I have decided to give her extra couple of marks because she stuck to the word count.*

This again gives an insight into his thinking at the stage of deriving an overall score from consideration of separate criteria. He clearly values conforming to the word count relatively highly since it seems to have a big influence on the final mark.

However, on another occasion, T4 Sam ignored the word count limit and he did not count it as a vital factor affecting the overall rating:

> *Well, not bad. We've got sixteen adjectives in there, which is good. His word count's probably bigger than Arikan. It's probably too big, the word count, but I won't penalise for that.*

In this quote, he also compared with a previous student regarding the word count as if he considered that student's script as a model (see 4.6 for more on this issue).

T6 John gives an example of arriving at an overall judgment overtly taking into account a wider range of criteria than the examples above. He refers to grammar, vocabulary and the text structure, although apparently weighted in favour of the first two in arriving at a score:

> *Yes, so no obvious conclusion there but, as I said, we don't specifically ask for that. But the grammar, not too many grammar mistakes, so mostly prepositional, a couple of plurals – again, these are quite common mistakes. But because there aren't that many – there's at least one grammar mistake in each paragraph, but because it's mainly prepositional and plurals it doesn't really make it that much more difficult to read. So I'd give that one a 7 out of*

> *10, mainly because there aren't that many mistakes, for grammar and vocabulary.*

Here T6 comments first that the conclusion was considered something not to be the focus for his evaluation as he said he did not ask for that. And then when he talked about grammar, he seemed to evaluate it in a positive way by saying not too many grammar mistakes, and that the types mentioned did not hinder comprehension by the reader. Vocabulary got only an incidental mention. Nevertheless, it is grammar and vocabulary that he implies justify the overall mark.

However, in another script, he had a different focus in his talk about justification. He commented:

> *Going through it again, now, for content, we clearly have an introduction. We have two paragraphs in the main body, which is good. And we have a conclusion but she didn't say it was a conclusion; she really needs to signpost that. The content, she gives some very good examples, albeit because of the grammar and vocabulary they're more difficult to understand. So out of ten I would give that a five.*

In this case, he spends more time talking about the text structure, then only briefly refers to content, grammar and vocabulary. However, with T6 it seems that the time spent talking about criteria does not relate directly to the weighting of the criterion in arriving at the overall rating. In both these instances he bases the overall score more on grammar and vocabulary, regardless of the time spent talking about them relative to other criteria.

Another example offered by T6 John cuts through the issue of how to weight, relative to each other, criteria such as those above related to specific levels of language. Instead he decides the overall score/rating based on the more over-arching communicative criterion of comprehensibility. Interestingly, however, he sees whether he himself can

understand the text as less important for deciding an overall rating than whether, for example, another non-native reader would understand it:

> *The reason I give it a four is because I can understand, as a native English reader – her vocabulary and grammar is good enough for me to understand what she's trying to say, so that's what gets it off the bottom. It's not a one or a two, or a three for that matter. But what stops it getting even halfway is the fact that there are so many mistakes, and a non-native reader would have to have a very good level of understanding, I think, to be able to read this and understand what's being said. So that's why I'd say four.*

Another general criterion we found used, that is not in principle linked to particular levels of language, is error frequency. A good example that illustrated this is T3 Zain,

> *Well, she got a problem with the punctuation mark there, but it is happening once, so it is not an error. It is just a mistake.*

Here, T3 Zain, commented that the student had a punctuation problem but he did not take it into consideration in arriving at an overall rating as it happened only once. He further shows some sophistication, possibly derived from previous training, in connecting frequency with the distinction between a mistake and an error. He appears to be reflecting a difference made by applied linguists in the field of error analysis, between a slip of the pen (or finger on the keys), due to momentary inattention or fatigue etc., termed a mistake, and a failure due to having not learned or mis-learned some language feature, termed an error (James, 1998). The first potentially could be self-corrected by the writer during editing, while the latter could not, since it reflects competence rather than performance. Naturally in most assessment the latter would be rated more serious than the first.

Often such decisions are accompanied by the teacher saying something that indicates why he/she made the decision he/she did. Hence, further examples of this will be

presented in the next section where the wider justifications offered for them are also considered.

4.6 Non-text-based support for rating or scoring judgements

Much of the TA and post-TA retrospective interview data consists of statements about core aspects of rating such as referring to criteria, often in either a positive or negative way (as exemplified in 4.5.2), and coded as either Language judgment or Rhetorical and ideational judgment (section 4.5.1). Another part of what we might see as the core activity of scoring/rating is reflected in talk about how ratings or scores are being combined to produce a final rating/score (4.5.3).

In addition to such core activities, however, there were several other identifiable kinds of mental activity reported. Particularly important is where teachers talk about why they rate or make decisions as they do, or indirectly evidence how they support those decisions. This is the focus of **RQ5.**

> **RQ 5: How do the teachers support, explain or justify their rating/scoring?**

Clearly the primary source used to justify and support ratings/scores is the text of the essay itself, which we do not consider further here, as it was widely represented in 4.5 and is described further in 4.9. It constitutes the default source of information to support rating, and is of course one of the three main sources when rating occurs in a test/exam context (along with the task prompt and the specification of criteria and scale). Rather what is interesting here is the range of other things which may be invoked by the teachers while arriving at a rating, such as we might expect from what was revealed in 4.3. We can see some examples relevant to this in the last column of table 4-9. They

are valuable in helping us understand the teachers' mental processes while they do this sort of rating of class practice essays.

4.6.1 Retrospective considerations

Apart from the text itself, an important source used in rating writing (as in exam assessment of writing) is obviously the task instructions. After reading some of an essay, T4 Sam for example says:

> *His structure's incorrect: he's made three paragraphs and there should be one. This is actually a really bad example of what was required here. Even if he has used lots of adjectives, and used them quite well, he hasn't actually achieved what we were asking him to do here. For that reason, I can't give him a good mark. I can't because he's not actually hit the target. By not making the target, well, let's see: let's go back to my rubric.*

Despite some good features of this essay, T4 Sam weights paragraphing high relative to the others because clearly the task instruction required one paragraph only. Correspondingly in another instance Sam is favourable for the same kind of reason:

> *So, this is actually right on. I am looking at this, you can see this is exactly the word count that she's been given. She's met this perfectly, you can see that.*

Other references however are often to the teacher's expectation about the writer, which implies (unlike a rater in exam conditions, usually) that they have identified the writer of the essay from their name on the script or other cues, and remember how that writer wrote before, and other features of the writer. This, on occasion, is clearly used to inform the rating or score that the teacher is in the process of arriving at, besides what he has read in the essay.

T5 Gena here for example is using information about the writer's past performance to support her overall favourable rating, although still giving the evidence of the essay itself the dominant role in the decision:

> *OK, so it's quite a good essay. I know this student and I know she is quite good. She's very good, actually. Her writing is consistently good, and this sort of confirms that. However, yes, there are some mistakes. Nothing major,*

T1 James put the point even more clearly:

> *I think when you're marking, you might compare it with what that student normally produces: is this good for Grazia, or is this a good effort compared with what Grazia normally produces*?

Other justifications sometimes arise from more speculative supposition about the thinking process of the individual writer. T5 Gena here, for example, seems to justify being especially negative about a missing part of the argument not just because it is missing, but also because she attributes that to the writer either just following his own feeling, rather than the task instruction, or his forgetfulness:

> *He just felt he'd written enough or maybe he forgot the question. That is, what's the negative for me. Other than that, it is an excellent essay.*

T1 James by contrast makes more explicit reference to his background knowledge of the class as a whole presumably relying on general memory of their performance on previous writing tasks, when he says:

> *He has not made as many mistakes as many of his classmates might have made. This is better than average.*

At this point the judgment is voiced in relative terms only as better than average, presumably meaning better than the average of the other students in the class, in norm referenced terms. Later James uses this to arrive at an overall absolute IELTS score for the writer of 5.5/6.

In another extract T2 Bob similarly refers to the class rather than the individual writer, but in terms of what he expects of them rather than what they typically achieve:

> *This one is marginally lower than I would expect overall, but generally of the level I would expect in the class. It's marginally lower simply because of the grammar. The other aspects are good enough for that level.*

Sam also refers to classmates but, unlike the above examples, in terms of their actual scripts written at the current time rather than his memory of their past performance or expected level:

> *So rather than just saying "he is intelligent" or "he is a good student", he has said things like "he's from a big family and he works in" – I don't know the name of this city – "at this university and he is single" – OK, that's an adjective – "but he is planning to get married". All right, which is really good. And nobody else has been able to do that, to use this complexity in sentences, to use an adjective and then a noun phrase afterwards, where within this noun phrase he's actually describing what his friend is like.*

Here the writer is rated highly not just on what he wrote, but by comparison with what others wrote.

T6 John exhibits the same use of comparison between essays written at the same time by different writers to guide the rating, but focuses just on two instances:

> *They've completely drifted away from the topic. And also, this one was strong in vocabulary whereas that one was strong in organisation and weaker in vocabulary, so the strengths and weaknesses are different.*

Perhaps the ultimate use of reference to what peer writers do is seen in T2 Bob who offered a unique reason for ignoring the error in one of his scripts:

> *"Facebook and Twitter". Nobody can spell "Twitter".*

Here, T2 Bob claimed that the word *Twitter* cannot be written correctly by everyone,

by which he presumably meant his students. Hence, he ignored the spelling mistake on this occasion.

Moving away from norm referenced comparisons with the student him/herself or other writers, T2 Bob on the other hand justifies his being hard on grammar in a different way by saying:

> *They know I am not happy if you have a sentence without a verb. This really irritates me.*

In this case, the reference is presumably to what the teacher says in class or said in previous essay feedback. The implication is that the students should be especially aware of such errors, and of the teacher's special dislike of them, and so avoid them, so he weights such errors very strongly when arriving at an overall score, as indicated by his reference to irritation.

T6 John more explicitly refers to his current instruction, which in turn is related to the students' current level:

> *I've asked for at least 150 words, and just going through it I can see that it's around about that amount. So, I'm not too worried about the word count, because at this stage of their development I'm getting them used to good grammar, good vocabulary and good organisation. The word count will become more important as we go through the course.*

Here he is clearly linking how he weights the criterion of word count to the point that the student is at in the current course.

T5 Gena again refers to her past teaching/feedback, showing additionally some irritation:

> *He is making mistakes with countable nouns which is a common mistake and I keep telling them, this student in particular.*

Finally, retrospective reference to what has been taught is not always used negatively:

> *Um, vocabulary? Well, we've got "affected", which is quite good; "different aspects of life" – just make sure then it's not in the question, no it isn't; "tend to", which is something we taught in the lesson, which is good; another "tend to", good; "have fun" is correct; "playing games", yes. And "generally" as well as "tend to", so this person has tried to apply what we've been doing in class. "Doing research" – correct collocation. Erm, "applications" is good. Um [...]. We've got "make friendships", which is wrong. Overall, though, I think it's pretty good vocabulary, so I'm going to mark that up.*

Here T2 Bob seems favourably impressed by the student trying to use what was taught.

4.6.2 Prospective considerations

Interestingly in some instances we find teachers using information of the same types as those above to inform and support their criteria and their weightings prospectively rather than to justify them retrospectively. That is to say that instead of providing justifications from non-text based sources during or after arriving at a rating, they refer to these sources at the start. It is as if they are activating information that they see as relevant to the rating activity before doing it.

Bob for example refers to the writing task prompt in this way:

> *What I bear in mind? Well, I always like to remind myself of the question.*

This shows that, from the start, he makes himself aware of the nature of the writing task.

T3Zain on the other hand refers to his memory of the writer's previous essays before starting, in order to guide his rating:

> *Well this is Dali, she writes some complex sentences in her previous compositions. So, let's see what she produces.*

Again, James refers to his background knowledge of the writer's L1 in this way:

> *When I mark for this particular student and because she is Italian, so they have complex sentences with multiple clauses in them which work in the same way as English. […] so those kinds of mistakes do not normally happen with the European language speakers, so I kind of know what to expect with people with different language L1s.*

There were however only very rare references to the writer's L1. This could be because even though the raters mostly knew what the L1 was of each writer, they did not in fact speak that L1 so L1 based expectations or explanations did not come to mind.

T4 Sam in another instance refers to using the comparison with the current set of compositions from different students, but this time before he has actually read the others:

> *I am going to revisit this to see if it is good as the other compositions or better.... when I read all the scripts.*

So also, T3 Zain derives prospective expectation from an inter-student comparison:

> *Well the first student was so grammatically accurate actually. I focused mainly on ideas and structures of paragraphs. ....., that student was kind of exceptional. So, other essays, it will be more comments on grammar, let's see.*

In another example T5 Gena makes a prospective evaluation by referring to what students have been told by her. As commonly in the retrospective instances, this is again associated with her personal attitude or affective reaction, i.e. irritation:

> *Well straight away this kind of makes me a little bit angry. I'm always telling them to write – because this is handwritten, obviously – to use double line spacing. They were given, I think in all cases except one who did the essay another day, they were given lined paper and they were asked – or they should know – to write leaving a space in between each line. He hasn't done that. They were given two pages; he's written on one side of one page and it's all closely jammed up in single line spacing. And straight away, I have to say, that makes me a little bit annoyed. And if I'm totally honest, I'm sort of feeling a bit negative already about the essay before I even start to read it. So I will tell*

> *him that, and I'll also tell him that he may get marked down a point for that, because he needs to learn that. It does annoy me. OK, let's have a look.*

In the following, T5 Gena again exhibits a high degree of the irritation which often seems to accompany reference to what the students have been told before:

> *A lot of spelling mistakes can be quite irritating, because sometimes, you know, you just stop understanding. You get to a point where you think "I can't continue with this; there are just too many spelling mistakes". Or other mistakes. Sometimes if there are so many mistakes that you just have to stop reading. But this was, even before I started reading, I just thought "aargh", because he's been told.*

However, in this case it is possible that her degree of irritation is not really justifiable by the fact that "he's been told". After all, however much one tells students to avoid spelling mistakes, a weak student may, with the best will in the world, be incapable of spotting all the instances where they may have made a spelling mistake and put it right (e.g. by use of a dictionary).

4.6.3 Consideration of rating impact on the writer

Finally, a rather different kind of justification for a rating decision is based on the rater taking into account what he/she believes to be the possible impact of the rating on the writer as he finalizes his judgment. This of course only applies where the rating is in fact going to form part of feedback to the writer, which was not always the case: some teachers on occasion distinguished between their own rating, shared only with the researcher, and what they would communicate in feedback to the writer. In those cases, any adjustment to the feedback to suit the learner did not impact on the rating itself, and those instances are described in 4.7.3.

T2 Bob illustrated this:

> *Yes, the basic problems are really affecting plurals and verb tenses. They're not all wrong. Some of them are right. Too many of them are wrong. Shall I mark him down one or shall I mark him down two? It's a bit cruel to mark him down two. I'll mark him down one and hope he gets the message. OK.*

Here, T2 Bob appeared to be working on one rating score which is to serve himself and the student. However, he is concerned about not appearing cruel, implying perhaps that this might demotivate the writer, while at the same time communicating a realistic assessment of the quality of the essay. Thus, his decision is a product not just of what is in the script itself.

T2 Bob in this next excerpt also shows that he is taking into consideration the effect of his rating as feedback on the affective side of the writer (*motivate*) as well as the cognitive/learning impact (*have the best effect*):

> *I try to be objective but equally I do include a subjective element especially with this group, I think what is going to motivate or what's going to have the best effect on the person who's receiving it. So if it's in between two categories, I will think what is going to be best for the student. Is it going to be best to encourage them? Is it going to be best to give them a little bit of a warning?*

T1 James, despite his IELTS examiner training, also seems on occasion to be allowing his care for the feelings of the student, and perhaps his own emotions, to affect the evaluation he is actually making and not just the version of it which he transmits to the student (as indicated by reference to student in the third person):

> *You know, you can say that in a positive way. Rather than saying "He's done this and this and this wrong", you could say "Well, he's only done this and this and this wrong". He hasn't made as many mistakes as many of his classmates might have made.*

In such instances the teacher's attitude, in the form of a desire to be positive for the student, appears to affect not only the feedback given to the student but also the rating actually recorded by the rater for his own purposes. In the context of practice writing done as course work, this may not be unusual, however.

T5 Gena illustrates a similar conflict in trying to award her summary rating score. She seems to be influenced by what the writer needs rather than what the essay deserves:

> *Possibly maybe a 5.5 actually – 5 or 5.5. I'm not sure if that would be enough for him if he wants to go on to study in a university. He's still got some work to do. It's a difficult one.*

To conclude this section, we summarise the possible sources of information that teachers can use in rating tasks such as ours in Table 4-10. It is worth noting that many of them would not be available to an examination rater, for whom the writers are unknown. Hence such a rater is largely forced to base the ratings just on what is in the script, together with the task specification and the rating scale and criteria to be used. It is the nature of the different rating task in our study that has produced this richer variety of information sources that the rater can draw on to support and provide justifications for rating decisions

Table 4-10 Information sources used by raters to support or justify ratings that they make, and for other purposes

Information**, available to be used repeatedly while performing rating / scoring** **Most of RIF-I and LF-I and some of SM-I codes cover some of these**	
Information sources	**Activity needed by rater in order to use the information source**
Rater background knowledge of writer in general (Prof level, L1, perceived needs, target level etc.)	Read name; Recall information from past memory
Rater background knowledge of writer's past writing	
Rater background knowledge of other writers in general, e.g. classmates of the writer being rated, and their past writing	Recall from past memory
Rater background knowledge of recent relevant teaching, nature of course, what point they are at in the course, and what the writer should know from that	
Rater beliefs about reasons for student performance	Recall from current memory
Rater's beliefs about impact of rating on writers	
Rating scales and criteria, rules for assessment etc. (whether provided or self-chosen, fixed or flexible)	Recall from memory; reading and interpreting available hardcopies
Writing task topic	
Writing task genre	
Writing task other specifications	
Script of writer being rated	Skimming, reading and rereading text; interpreting it; spotting things; editing or adding notes to the script to assist rating; recent memory of script read
Scripts of other writers being rated at the same time (in SM-J)	Recent memory; notes made

4.7 Wider aspects talked about

Our data was very rich and varied, and resisted coding into systems largely developed in the context of rating studies done in simulated writing exam conditions rather than real pedagogical writing practice conditions. The following are some interesting categories of report found in the TA and/or follow up interviews, other than those reported on so far. Although they draw on the same range of information sources as considered above (Table 4-10), they arguably lie outside a strict focus on rating/scoring,

since they do not very obviously assist or explain the ratings that teachers arrive at. At the same time, they throw important light on the way teachers doing our sort of rating of classroom practice essays elaborate on their rating while performing the rating process, so as to invest it with more meaning.

RQ6: What other kinds of comments do the teachers make, beyond those related to achieving the rating itself?

4.7.1 Suggested reasons for student performance

Some remarks were in effect suggestions about why the writer might have written what he/she did, or get the rating he/she did, especially negative features. Similar examples which appear to have actually affected the rating were discussed in 4.6.

Some were based on writer L1, such as this from T2 Bob:

> *Yeah, I know why is this? I think it is written in their first language. It is different.*

T3 Zain further uses his background knowledge of the writer, by identifying precisely what the L1 is:

> *I am trying to figure out what she means. Who is this? It's Joan. Yeah as I mentioned she is Kazak. Sometimes they write in their first language style, I am not sure. I have to reread in order to understand.*

It is not clear in many cases such as these that the interpretation of the reason for the error actually affected the teacher's rating of the error (cf. 4.6). We interpret them, as clearly in the second example, as designed to serve more as comments on problems reading the text rather than comments on the rating itself. Understanding what the writer intended to convey is, however, often a prerequisite for identifying errors and applying evaluation criteria, so to that extent has an indirect effect on rating.

Again, James says of a student who he has already given 5.5 to, based on what he wrote:

> *So, bearing in mind he's only a B1 level student…. so, to get a 5.5 or a 6 is a very good score for one of those guys. This is a very good essay for somebody in this class. This is better than average, yes. It's a good job.*

Again, he appears to be referring to the writer's level not to contribute to deciding on the score (cf. 4.6), but to add meaning to the score already decided.

Bob again offers two other kinds of possible reason (underlined); again, both rely on the teacher's background knowledge of the writer:

> *OK, he's thought about it, he's attempted to use logical sequence of his ideas. It's not all bad, but it is weak; it is weak. And I know, as I said, this student – he tends to write as he speaks. He's got quite good fluency speaking, quite a good level, but he's lacking academic writing skills and he's still got a lot to learn. And it's partly because he hasn't done the written work this term.*

Once again, we cannot say for certain whether these purported reasons had any impact on the rating, but it does not appear so. The purpose seems to be more to explain the rating rather than to contribute to forming it.

Another example from T4 Sam not only identifies why the writer wrote well (because he asks questions in class) but further elaborates in detail about the effect of this on other learners in the same class. In this way, the rater is treating the rating task as an occasion to reflect on teaching experiences and issues which go way beyond what is simply required in order to rate the quality of the script. In short, he is behaving like a teacher rather than just a rater:

> *I know this student, and this is probably his best piece of writing. So, this is really interesting. Because this student he is a kind of student who asks a lot of questions which is really good. And he will ask a question until he really understands what you are talking about. Not all the students will do that at this level. His friends at first were becoming a bit angry with him in the class*

> *but then they started to realise he was asking the questions they wanted to ask as well.*

These kinds of implications are further illustrated in the next subsection.

4.7.2 Implications for teachers and pedagogy

Some teacher comments which depart somewhat from evaluation were giving implications of some sort from the rating which had just been made. In a number of instances the implication is for the teacher's belief about the level of writer, which he/she seems to have recalled after rating the essay (if not before, see 4.6), and either confirms or upgrades, based on the rating he has arrived at for the essay.

For example, T3 Zain says

> *I am surprised because one of the teachers mentioned that he is bad in listening and speaking*

Clearly updating his knowledge about the student. So also, T2 Bob:

> *Better than I might have expected from him to be honest, Amjad.*

Also, T1 James:

> *Really impressed by Ali as he recently joined the class, he is an intermediate student not high.*

T3 Zain on the other hand shows that the essay he just read confirms his view that the student is of a lower level than that which he has been assigned by a placement test administered by the ELLP:

> *I keep saying this student should not be in C1B which is the advanced level. I would put him in intermediate or upper intermediate.*

Other references are to what can be inferred about what has been learnt from teaching. T4 Sam says:

> *He actually took on board what we have been teaching over the 5 weeks.*

Again, T3 Zain registers surprise that the writer of one of his scripts shows signs of having learned something which in previous classes he seemed not to have learned:

> *He is fairly clever. He is using this. I was telling him this in the class. He definitely seemed not knew this. But he does.*

Sometimes the implication is for the teacher's feelings rather than his/her background knowledge of the student. This may take the form of surprise as in the previous example, or more often disappointment. There are many emotional reactions (code self-monitoring B 01 d) which consist of the words *Oh dear!* and other emotional reactions associated with negative judgment such as from T2 Bob *I am really disappointed in this from John* and from T5 Gena, *this makes me a bit angry*. There are however also some positive emotional examples, such as from T5 Gena: .... *it's a nice complex sentence, so very pleasant surprise*.

Emotional effect is clearly stated by T4 Sam:

> *He hasn't made many spelling mistakes and you know what? His grammar is pretty good. But, I like to say he's using these more complex noun phrases to describe his friend rather than just single or paired adjectives. It's really encouraging actually, really encouraging. Because it makes me feel good as a teacher.*

Returning to cognitive rather than affective effects, instruction related references most often occur in relation to future learning and teaching, as illustrated by T6 John:

> *So, overall, it's good but could be better, which is why it's got that twelve out of twenty. And again, I think 60 per cent at this level is a good mark. One thing*

> *that's worth bearing in mind is that, if this had been an exam, they need around about 70 per cent to be able to move up a level. So, they're getting there, and we have another ten weeks to work on this, so I would fully expect this second one to be able to move up after another ten weeks of instruction.*

On another occasion the same teacher explains how he will exploit the specific information from the current essay rating to guide his teaching of this student in the future, using it, in effect, diagnostically. This of course can only happen because he rates on detailed criteria in an analytic way:

> *The interesting thing about IELTS, of course, is they give different scores for reading, writing, speaking and listening. Because this is a writing task, I'd say it's towards the lower end of the B1 level. Certainly, she's in the right class, the right level, but it's towards the lower end. So, a bit of work needed doing on that, but I've now got enough information to be able to correct her when I see her next, and show her where the improvements can be made. OK.*

Another interesting example is when T5 Gena expressed her views about the ability of one of her students, but in a way that presents him not as being in a state of deficiency but rather as in the process of learning.

> *I think perhaps it's his case that he can't see exactly what he needs to see because there's so much … he's not seeing clearly. So, he's not really a very competent user yet. But he's getting there; he's getting there.*

These last excerpts demonstrate that the teachers, regardless of their experience, training, and cultural background, conceptualize their students as located in a developmental process of learning. This finding corresponds with Erdosy's (2000) study, which found that all his rater informants were showing the same attitude: "each of our four participants provided that link [a link between performance and proficiency] by constructing a developmental trajectory for presumably adult learners of a second language, on which they could place the writers of the compositions" (2000, p. 106). As Erdosy also found, "[These] profiles could then be related to students' performance

on tests, facilitating the estimation of their potential from their current position on a developmental trajectory". (2000, p.36). T6 John and T5 Gena in the excerpts above are clearly envisaging their writers as being on just such a trajectory.

Such comments by teachers evidence the extent of their contact with students in a wide range of situations in and beyond the classroom which enabled them to construct a rounded picture of students' development and abilities. Such comments are not about the rating process itself at all, but about its implications and consequences for learning and teaching. However, they do illustrate how, in the sort of rating task that our teachers were performing, such comments get intermingled with the rating process. In the assessment of writing in examinations/tests this sort of comment is again unlikely because the rater usually does not know the identity of the writer, or even their course, and in any case, is solely focused on rating itself for the test purposes. Hence such comments are not found reported in most studies of the assessment of writing. In our study, however, rating is occurring more as part of the teaching than testing process, with writers known to the rater and being given individual feedback in addition to rating. Hence such comments are natural. As we suggested in 4.3, the teachers see themselves as teachers more than raters/assessors, so this symbiosis is to be expected.

4.7.3 Feedback to the writer

Most teachers talked about what feedback they would or would not give to the writer, or gave such feedback as part of the TA. Thus, they in effect elaborated on implications of the rating for the writer.

T1 James stood out by wording his TA protocol for some essays not just in the form of voicing his thoughts, but also in the form of voicing a commentary on the feedback that he was writing on the essay script for the writer:

> *This is not a bad introduction. It's very short, but "things" is a bit rough. So rather than "things", I think he should probably say "aspects", perhaps. So, I'm just going to correct that and write "aspects". "Of eating junk food regarding health and in economic terms": that's fine. It's short, but it's OK. So, I'm just going to put a tick and "Good" for the introduction. OK, "One the one hand" – he means "On the one hand", so I'm crossing out "e"*

More often, however, the teachers voiced feedback later, after deciding on the overall score or rating to award, for the purposes of the assessment task. T1 James provides a straight example of what feedback he is writing on a script at this point:

> *So, the biggest problem Hakim has here is that there's a lot of irrelevant stuff, the introduction is too short and there's no conclusion; so that's what I'm going to write. So, I'm going to write, "Not bad writing, Hakim, but you have forgotten about the question in two paragraphs, and have wandered off the topic. Your introduction is too short and there is no conclusion. You need to focus 100 per cent on the question and to organise your paragraphs better."*

Here the teacher sees it as part and parcel of the rating task that he is performing for the researcher that he also gives feedback to the writer. Although he seems to see the rating and the feedback as distinct activities, in the example above the two do not really diverge in content. His rating of the script and what he tells the writer for pedagogical purposes as feedback match.

Elsewhere, however, T1 James for example seems to suggest that he might make a difference between his actual rating and what he 'writes' for the writer to read. In doing this he makes full use of his background knowledge of the writers, and his understanding of the impact of rating on the writer (cf. 4.6):

> *I care about my students, and so I'm pleased to see, "Oh, look, Ali did that. That's good." He's new in the class this week. I didn't expect him to do so well. It might also affect what I would write at times, because obviously, I expect more from some students than others. And so, if a student does better than I expect, then I might be more positive about what they've written;*

> *whereas if another student who's stronger writes exactly the same essay, I would obviously be positive, but I wouldn't be so enthusiastic because I would say, "Yes, I expect this from you, Anwar. This is no big deal." OK? So yes, partly there's kind of an emotional reaction – you're pleased when a lower-level student produces something better. But partly also for the student's benefit, I want to encourage a student whose writing is not normally so good; so, if they've done something that's better than I expect, then I want to let them know, "Hey, well done, Ali. You're making progress." That's important.*

Here the teacher shows special awareness of a need to take care of the affective side of the writer: he does this by separating his real judgment from the feedback rather than keeping them the same but simply altering his judgment (cf. 4.6.3). This means he will not necessarily give two writers who produce the same level of writing the same feedback.

In the above examples the teacher tends to be more positive in the feedback than he/she believes is perhaps merited by the student's true performance. Interestingly T3 Zain illustrates the reverse, because he is thinking of how the peers of the writer might react:

> *I will try to avoid words like good and very good because they are sensitive if they see their colleagues got 'very good' and 'good'.*

A related issue concerned whether or not actually to give the overall score at all (if there was one) in feedback to the writer. T3 Zain is very clear about not giving marks:

> *I am not giving them marks actually, but if I am marking this for IELTS, I would give it 7.5, definitely 7.5. However, this is just for you. I am not writing it on the sheet.*

He attributed this decision to being keen that the students will not focus on the awarded score and neglect the other feedback which is more important.

T2 Bob also shares the same view in that he prefers not write a score. Instead he prefers to encourage them to meet their target. He stated:

> *The students often ask me about the score and I try to get round it a bit. I say to them "what's your target?" and then I make a comment in relation to their target, such as "well, if that's your target you're going to have to work a little bit more to reach that" or "you're going to have to work a lot more to reach that" or, you know, "you may be near where you want to be". But I won't actually give scores.*

Overall it is clear that decisions on feedback mostly are made after those made in the process of arriving at a rating/score for the writing itself, and are kept separate. However, there are signs that, for some teachers on some occasions, considerations of what would be suitable feedback may wash back into the rating process itself (as described in 4.6.3). As we have noted in other sections above, these issues are unlikely to occur in test/examination rating of writing where the rater does not know the writer and has no brief in any case to supply feedback to the writer, other than the summary score which should only reflect the merit of the essay script.

4.7.4 Personal reaction to ideas or to students

During the evaluation process, most of the teachers on occasion expressed their personal attitude to the ideas or content in an essay. Regarding the personal attitude toward the students' ideas, the best example that illustrated this issue is T3 Zain who showed his ideas clearly and frankly here:

> *"Charles Darwin". Ok, Uhha "said that the labour". Well, why the labour. "Said that labour has an essential role in transformation process from ape to man". Well, that is a bit silly. She still believes in Darwin.*

Here, T3 Zain could not hide his feeling about the student's idea. He even called it a silly idea. Moreover, his stupefaction continued by denying that Darwin's idea is correct.

In another example, again T3 Zain said:

> *"as nature created human beings". SubHaan Allah [Allah almighty], those Kazak people are so crazy. "As nature created human beings". It sounds they do not believe in God. Anyway, "as nature created human beings", "nature created human beings".*

Here, the teacher, who is an Arab and we assume is Muslim, reflects his own beliefs in reaction to the non-religious view of creation reflected in the essay. There is no indication, however, that this reaction contributed to his summary evaluation of the essay. In such instances the teacher is responding to the essay communicatively, as a real reader would, rather than as a teacher or rater.

In another example, T3 Zain, after querying a language point (*engaging*), challenges the writer's support for children working:

> *"Children must engaging in some paid work". Why engaging? Right, she seems to be with child work. Right. That is weird.*

Here, T3 Zain, disapproved of the idea that was offered by the student writer by saying it is 'weird'.

Another way of expressing personal reaction to student's ideas is to laugh. In the next extract T4 Sam was laughing out loud about an idea found in one of his student's scripts.

> "*She tells me she likes to have knowledge about boys. She wants to find an optimistic boyfriend". Lol… she is lol…oh well she is describing her friend though.*

The laughing response was rarely found among the informants but in this situation particularly, T4 Sam could not stop laughing when he read this part of the composition. However, his laughing was in a positive way not sarcastic.

Our observation that none of these personal reactions seemed to have an effect on final rater judgements of the overall writing quality is quite different from that reported by

Sakyi (2000) for one of his participants, whose assessment was based on his own biases concerning the ideas expressed:

> "…the language is reasonably acceptable but I certainly couldn't give it a 4. It's even allowing for bias, it's the kind of stuff that sister Mary Theresa in a small Ontario school would encourage their student to write and it has all the tiresome, excessive […] but certainly doesn't go nearer to God to me, …ok, so I have admitted my bias […] I think it is brainless. I do not think it has any thought to it. It is reasonably clearly written and I would give it a 3, 3+ but no more than that and not have a bad conscience about it. So, this would definitely be a 3". (Sakyi, 2000; p. 136).

4.8 Sequences and stages of rating related activity

In the above account, we have occasionally referred to teachers doing or talking about things in a certain order. However, for the most part we have considered the kinds of activities they engage in, information sources that they access, and comments they make, without concern for sequence of events. Indeed, our entire coding scheme, like most coding schemes, does not in itself capture anything about the order in which things occur, simply what kinds of things occur? In this section, we therefore attempt to fill this gap and answer **RQ7** through painstaking analysis of the TA protocols to try to find overall sequential patterns.

> **RQ7: What common sequence of activity can we find in the rating behaviour of our teachers?**

With respect to the order of essay scripts, most teachers simply started reading and assessing the compositions in the sequence they appeared in. This outcome corresponds with Cumming et al. (2002). The exception was T4 Sam whose habit was to read the

entire set of scripts to be marked before starting on the detailed assessment of each one. In fact, unlike any of the other teachers, he had already performed this first reading of the scripts before he came to our TA session. While this was not what we intended because it robbed us of the chance to hear his TA of the first reading of each text, we nevertheless continued with this participant because we had such a small number of participants.

The sort of sequence that is of greater interest, however, and which is the focus of our account, is that within the treatment of individual essay scripts. The activities we mean to cover include all those described in earlier sections, including use of criteria to evaluate, arriving at an overall judgment through decision making, using various information sources, and so on.

From scanning all the TA reports of the teachers made during each separate script, we identified a typical general sequence of activities that appeared in the TA transcriptions for each script (see Figure 4-1). The broad sequence of the process is presented in terms of macro strategic behaviours that are used among our teachers, such as have been mentioned in Cumming et al.'s study (2002). These strategies, according to Cumming et al. (ibid), account for "behaviours that the [teachers] exhibit when deciding how to deal with the overall set of compositions" (p. 72). This process was natural as we had not directed our informants about this behaviour because our aim was to discover how informants approached the task independently. We found that a three step model was appropriate, similar to Cumming's et al., (2002) model, and indeed broadly found in a number of models which we reviewed in chapter 2. With each of the three steps, however, we added more information which is seen in the second row at each stage.

It cannot be claimed that the framework of the teachers' practises that we present here exhibits any permanent or universal sequence of the thinking processes that the teachers engage in to evaluate their students' compositions. Nevertheless, it captures the broad sequence of the events which constitutes useful starting point.

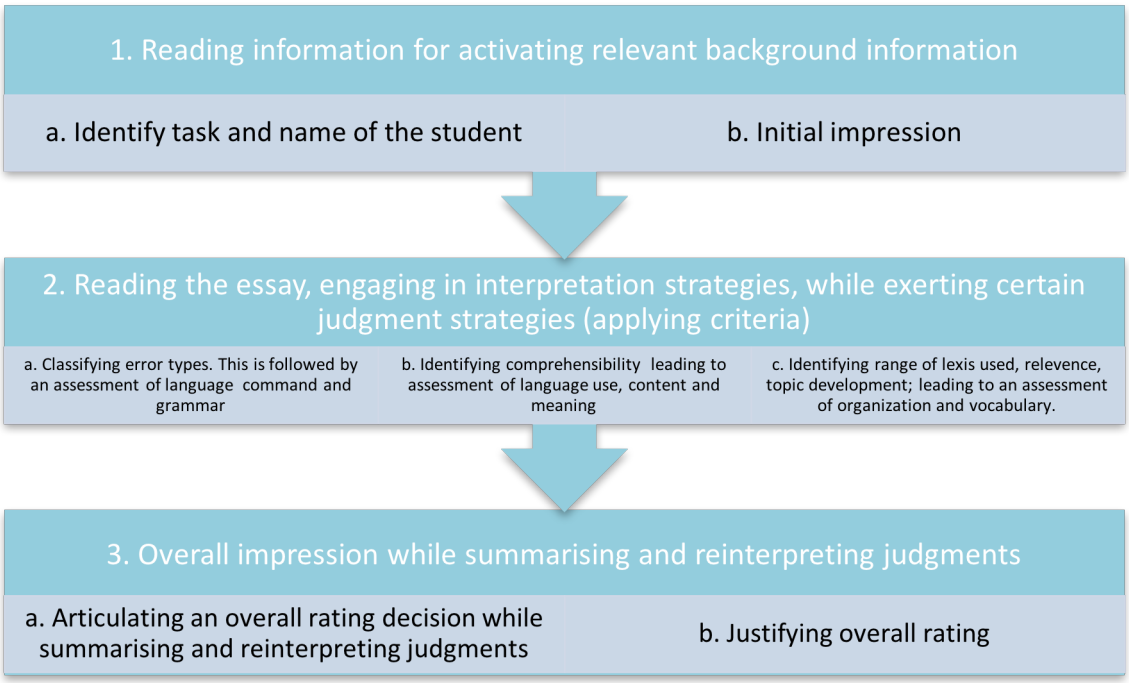

Figure 4-1 Sequence of the rating process while evaluating students' compositions.

It is of importance to highlight that the categories listed side by side at the second stage have no fixed arrangement among them.

4.8.1 Stage 1: Reading information for activating the relevant background information

We here look in more detail at the activities reported in the first stage of our model, associated with the first reading of the student's name, the task prompt, etc. Almost everything done at this point can and does create expectations which overtly or covertly may influence the rating that occurs at later stages when actually reading the text (cf. 4.6.2). The rater's thought processes at Stage 1 can quite likely guide the overall judgment leaving little trace even in the think aloud data.

Also, it is worth pointing out that in other studies (e.g., Lumley, 2005) of the type which is focused on assessing writing in the exam situation, at this stage a rater might not have access to the student's name but instead might for example reread the rubric for the scoring or rating system provided. That of course did not occur in our study since no such system was imposed.

Sometimes stage 1 was so short as almost to be non-existent, as in this example from Bob, though often it was more extensive:

> *Right, I must just remember the question. The question is about younger and older people using technology: do they use it in different ways? OK, let's look at Amjad. OK, spelling [...].*

Here stage 1 consists only of a quick review of the writing task prompt, before proceeding to stage 2.

4.8.1.1 Identifying task prompt

During the first reading, almost all teachers, except for one teacher T4 Sam, read the task prompt for many reasons (Table 4-11). Firstly, this occurred for the benefit of the researcher. Secondly, it was driven by the rater's own need, as stated by one informant:

> *I need to read the task because it is easy to forget it.*

This does not mean that the writing task prompt was not also read again at later stages in some cases.

It was clear that at this stage a certain kind of expectation, based on the nature of the writing task, was reported by some teachers. Often, these expectations influenced the initial overall impression and may have affected the overall evaluation. T3 Zain provides an example of such an expectation:

What was the task, the task was, um "in many countries", that is the second task of ELTS, "children are engaged in some kind of paid work, some people regard this as completely wrong, while others consider it as work experience" .... What I expect from the students is to write two paragraphs. One with and one against and one their opinion and conclusion and introduction. So probably five paragraphs. One, two, three, four paragraph. Let's see how that went.

Table 4-11 Frequency of reading task prompt

codes	Feature	T1 James	T2 Bob	T3 Zain	T4 Sam	T5 Gena	T6 John	T
01	Read task prompt	8	6	6	0	2	2	24

Table 4-11 shows that T4 Sam had no instance of reading out the task prompt in the TA. However, in the course of his evaluation he did refer to reading it. This demonstrates the limitation TA data, specifically the well-known limitation that it is incomplete.

4.8.1.2 Identifying student's name

During this stage, the teacher often first read the name of the student and on some occasions this process was associated with some comment about the student and his/her level or background (cf. section 4.6). In Table 4-12we illustrate some of these comments.

Table 4-12 Comments about student's level or background at the first stage

Number	Name	Comment about student's level or background
T2	Bob	*Now, Hakim, I don't expect this to be all that good. Hakim's grammar is not great. He's a great speaker but he's not a great writer, so let's see what he's done*
T3	Zain	*Let's start with number 1, she is a Saudi lady from C1B.*
T 3	Zain	*In many countries children are engaged" … ok that's the question. Ok who is this? Oh this is (Pazat). She is nice. She got 7.5 in listening. So, let's see how she writes??*
T1	James	*Who is that? This is Barbra from Turkey. Umm, as far as I know she got problem in grammar I think. So let me start.*
T3	Gena	*Ok that is the Greek, I think his English is not really bad but his handwriting is so difficult*
T4	Sam	*We continue this read-aloud marking of these paragraphs. So Yuona has done well here. Let's see.*
T5	Gena	*Right, the third one. OK. This one the handwriting, it's very neat but it's always a little bit difficult to read her handwriting. So again, I know who the student is.*
T6	John	*This is Doura this lady is an Arabic reader and writer, some words are back to front, or some letters are as well. So it's the sort of things I'd expect to find. Let's see.*

It should be noted that T4Sam's quote bears witness to the fact that, as we indicated above, he had in fact already read the script before coming to the TA session, and clearly had already formed an initial evaluation from that. Hence his first stage with us was possibly not really the first of the three stages in our sequential model but possibly part of the second.

Some teachers read the student's name without further comment, but even when that happened, we may surmise that, where the teacher knew the student, some activation of whatever mental schema he/she possessed for that student occurred. Furthermore, although some teachers did not mention the student's name at Stage 1, at later stages during their evaluation process, when they fully engaged with assessment process in

the middle, there was always a reference to the student's name, level and sometimes nationality. Some participants therefore only appear to be an exception to the generalization that the raters all engage in similar activities at stage 1, even if they appear with many essays to go straight to reading and commenting on the essay without mention of anything except perhaps the title. However, it is often clear that they have also tacitly registered the student's name and background, without this appearing in the TA report. This can be seen for example with T1 James: we can deduce this in his report on essay two from the fact that although he makes no mention of accessing the name at the start, he refers to the writer as *she* all through, and by name at the very end as *Grazia*.

The activated knowledge about the writers sometimes guided the path of the evaluation and sometimes had an impact on the rating all through. In support of this, Table 4-13 shows some instances of correspondence between what was said at Stage 1 and at Stage 3. T3 Zain for example at the start of his evaluation of an essay by Dali seems to recall a positive feature, that she writes complex sentences. Interestingly however by the end, when this is referred to again, it has become a negative trait, associated with lack of clarity.

Table 4-13 The effect of prior knowledge of the writer's name and background level in the process of evaluation

Name	Stage 1	Stage 3
T3 Zain	*Well this is Dali, she writes some complex sentences in her previous compositions. So, let's see what she produces*	*Well yes she writes very well actually with few grammar mistakes. urr but she is a bit ,, as I mentioned earlier ,,well she is a bit passionate in the way she writes and abstract and sometimes you know that makes it hard to follow. Starting thinking about the sentence level and you forget about the paragraph level.*
T5 Gena	*OK. Number five. OK. Oh, right. Yes I know him, is a little bit messy sometimes, but let's see. […] Well I know this student and he's done very little writing during this course, … He hasn't got much experience of writing this type of essay,*	*OK, he's thought about it, he's attempted to use logical sequence of his ideas. It's not all bad, but it is weak; it is weak. And I know, as I said, this student – he tends to write as he speaks. He's got quite good fluency speaking, quite a good level, but he's lacking academic writing skills and he's still got a lot to learn. And it's partly because he hasn't done the written work this term. So it's not all bad, but it's not great.*
T4 Sam	*[…] he's an Arab student and he struggled. When he first came into the class, he really struggled with his writing. He's a good speaker.*	*Wow here he's done actually very good. He kind of concentrated and done exactly what he needed to do, but done it in a more skilful way than some of the other students, which is really, really promising. So that goes against the stereotype that we tend to think about an Arab student as a writer.*

T5 Gena's example in Table 4-13 illustrates an interesting point which is that Stage 1 and Stage 2 may overlap. At the very beginning of her TA about the script she activates background knowledge of the student (Stage 1) but then she starts reading the first paragraph of the essay (Stage 2) and her exploitation of her background knowledge of the student reappears again at the end of that paragraph. T4 Sam's example of Stage 1 again evidences the fact that he had already skimmed the scripts, so his stage 1 is not really totally initial.

T5 Gena in retrospective interview shows explicit awareness of the possible conflict between her knowledge of the writer and the evidence of the essay. She puts it down to an inevitability of human nature (cf. 4.6.2):

> *I was just saying it is difficult sometimes. You know, we're only human, and sometimes when you see an essay and you know who wrote it, you kind of already come with expectations, and possibly sometimes it could be unfair on the student because you're judging them on their past work as well as what you're judging now. I try not to do it but, as I say, it's human nature, I think.*

But she believes despite knowing the student's name, it would not affect her evaluation when she clearly stated that:

> *I do not think it will affect my criteria.*

Therefore, she continues to clearly state her expectations of particular students when she starts her evaluation process:

> *My expectations, obviously, are, as I said, I see the name and I'm thinking, oh. But the last one, I knew that was going to be quite good. So, you kind of – you have where you think it's going to be, but sometimes they surprise you. And that's always nice, if it's a good surprise.*

4.8.1.3 Initial impression

At this stage, it is clearly evident that an initial impression is developed, after the teacher:

- Identified the task
- Identified the student's name and the writer's background
- Initially encountered the text

By this point the content of the composition is broadly understood and surface features such as physical layout and surface errors which might attract the teachers' attention

have been noticed. The teachers focus in this stage on building up an impression of the compositions. The impression that usually builds up in this stage is mostly not a fixed impression that can affect the overall judgment.

T5 Gena provides a nice example of where the initial impression from an essay, based on a very superficial feature, was negative, especially when she also drew on her memory of what the writer had been told before (cf. 4.6). This might have coloured her entire later rating.

> *Hmm, well straight away this kind of makes me a little bit angry. I'm always telling them to write – because this is handwritten, obviously – to use double line spacing. They were given, I think in all cases except one who did the essay another day, they were given lined paper and they were asked – or they should know – to write leaving a space in between each line. He hasn't done that. They were given two pages; he's written on one side of one page and it's all closely jammed up in single line spacing. And straight away, I have to say, that makes me a little bit annoyed. And if I'm totally honest, I'm sort of feeling a bit negative already about the essay before I even start to read it. So, I will tell him that, and I'll also tell him that he may get marked down a point for that, because he needs to learn that. It does annoy me. OK, let's have a look.*

In fact, in this case the final rating at Stage 3 recognised that the essay was in fact good despite the bad impression created at the start:

> *But the rest of the essay is quite good. Again, I would say that's one of the stronger ones. So, I would give him 8 out of 10, despite the fact that he didn't use double line spacing. I think he's written well.*

At the first Stage the teachers never assign a fixed score or state their final evaluative judgments. Rather, this decision of arriving at and verbalising the score can be found throughout the process of evaluation in the next stages.

4.8.2 Stage 2: Reading the essay, engaging in interpretation strategies, while exerting certain judgment strategies (applying criteria)

In this stage, the teachers were engaged fully with the texts and the real evaluation process started. However, the teachers were different in dealing with the text features and their major differences in style at this stage will be described in 4.9. Evaluation comments made at this stage had to be stored in their mind to access at later stages. The participants conducted their decision making with complex interactive episodes of information gathering and judging.

4.8.2.1 Reading behaviour

After the identification of the task prompt and the composition writer's name, all teachers read their students' essays through one or more times within what we regard as Stage 2. The teachers did this without exception. In my coding scheme, under Self-monitoring interpretation, text reading behaviour is a main code with four sub-codes (Table 4-14).

Code 04 which represents reading the whole text without interruption was a relatively uncommon behaviour among our teachers. Reading and rereading parts of the text, followed by some comment, was the norm.

The following example from T3 Zain illustrates a common pattern. The focus here was the composition itself, not the task description and the student's name. It is clear from the extract that he reads the composition in fairly large chunks (in inverted commas). The comments during his reading were brief. However, they included primarily evaluative comments (coded as judgment in the coding scheme) like *That is a good one*, and comments showing the teacher working on understanding the text (coded as

interpretation), like *So that is the introduction*. The reference to being not sure about *fruition* was initially unclear: it might be querying the correctness (judgment) of the word or registering that the teacher could not understand (interpretation). It later emerged that the teacher was not familiar with this word, so counted as the latter. Notably, however, in all this there is no reference to an intended score or overall summary rating.

TU	T3 Zain N1.
1.	"*Is trending in many societies xxx in some paid work"*
2.	*Om, uh. Right, so that is the introduction.*
3.	*"Xx" Right, um. I am not sure about this one, "it's trending in many societies".*
4.	*So, I just underline that one. I can check that, ok.*
5.	*"In some paid work xxx". Right*
6.	*"in recent years, it has been increasing argument whether children engagement in such works can possibly x a success"*
7.	*That is a good one.*
8.	*"In recent years, it has been increasing argument whether children engagement in such works can possibly (be held) a success, that would come to fruition".*
9.	*Fruition, I am not sure about this word.*
10.	*"in regard to helping them build their personalities or not"*

Table 4-14 shows the frequency of each teacher's reading behaviour in their reading of the text, during and after which evaluation occurred.

Table 4-14 Frequency of reading behaviour during Stage 2 reading.

codes	Feature	T1James	T2Bob	T3Zain	T4Sam	T5Gena	T6John	T
02	Reading part of the text	42	25	87	25	47	39	265
Average per text		7	4.2	12.4	3.7	6.7	6.5	
04	Read whole text	0	0	0	7	0	0	7
05	Scan whole text	0	1	1	3	0	2	7
03	Reread part	7	10	19	12	3	10	61
No of texts analysed		6	6	7	7	7	6	39

As Table 4-14 shows, code 02, read part of the text, which is in fact usually followed by one or more interpretative or evaluative comments, then continuing the reading, was the most frequent analysis category found in the data with 265 occurrences in total. It is clear that at this stage of the rating process, the most typical behaviour is to read each essay in a number of sections or chunks rather than as a whole. T3 Zain does this the most, therefore implying that he reads in shorter chunks. Notably he is also the most frequent re-reader (code 03). This could be related to the fact that he is the non-native speaker among the teachers, so finds it harder to retain longer stretches of essay text in mind while evaluating them, and also for the same reason needs to reread more often in order to understand the essay and to pick up the thread of the text following a series of comments. In fact, he is also found to occasionally refer to not understanding what he is reading and to needing to refer to a dictionary. For example:

> *"Fruition" I might check that word in dictionary. I have never heard of it before. I suspect that "fruition" its sound English to me… <He then checks a dictionary and reads the definition>*

T4 Sam by contrast divided each script into only 3 or 4 parts to read. Part of the reason for this is that the writing tasks which he set were descriptive paragraphs, not full

argumentative essays in the style of IELTS writing tasks, which the other teachers set. Hence Sam's scripts were shorter – with a target length of about 80 words rather than 200 (see chapter 3). For the same reason, perhaps, T4Sam was the only one to report reading the text as a whole without stopping, which, as we have described earlier, he unfortunately did before he came to the think aloud session, contrary to our instructions.

Overall the figures in Table 4-14 shows some tendency for those who read more parts of the text to also reread more parts (r=.547). Without T3Zain's extreme score, however, the correlation was strongly negative (r=-.849). The NS teachers who read more parts reread parts less. This could be due to an effect that reading fewer parts means reading longer parts. Longer parts are harder to hold in mind even for NS while then engaging in interpretation, and using evaluative criteria (judgment). Hence when fewer, longer parts have been read, the need to reread them is more likely to arise. This pattern is however not seen in Lumley's corresponding table for his data (Table 4.4 in Lumley, 2005: p150).

4.8.2.2 Interpretation and judgment

As our coding framework reflects, a great deal of what the teacher raters reported doing based on the essay text itself falls either into evaluative/judgment behaviours/strategies (fully described in Table 4-6), or non-evaluative/interpretation ones. As stated in the process model above, both occurred at stage 2, interspersed with the reading of the script.

As T5Gena shows in this excerpt, interpretation and judgment are often intimately intertwined.

> *.... and also, he starts by saying "they", but who's he talking about? He's using the pronoun "they". I suppose he's talking about students of fifty years*

> *ago, but he mentioned in his last paragraph that – "students of fifty years ago", he should have repeated that. He should have said "students" again.*

In this case, the rater has not decided from the start that there is an error (negative judgment). Her question could be seen as initially just a comprehension question by the rater as reader, trying to understand a text, so a matter of interpretation. However, the rater then follows through a line of reasoning which clearly ends up with a negative evaluation/judgment.

In the following example, from T2 Bob, the rater first questions the text again as if he is uncertain momentarily about his own understanding of it, or perhaps uncertain of his judgment that it is wrong. This we count as interpretation. He then corrects it, showing that he has in fact judged there to be an error. He then provides an overtly evaluative statement (judgment) that is interesting because it is positive. This illustrates a fascinating point which our coding scheme, despite its detail, did not capture. Although the rater has clearly decided that there is an error, which he corrects, he gives a positive evaluation. This seems to be because his evaluation is in fact not of what the student wrote but of the text after he has corrected it. He is, in effect, rating his correction as *nice*, rather than the writer's script. We have not noticed this phenomenon referred to in the literature.

> *"Spend a long time on technology they will lose their time". Mmm. Well, "lose their time"? That's not English, is it? – "they will waste their time and energy without any benefits". That's nice and good*

Since evaluation criteria have been extensively illustrated already in earlier sections (especially 4.5), we illustrate further here only some of the more frequent interpretation categories found in our data especially at the second stage of the process.

Many interpretation categories in our coding system focused in effect on discerning the message or content that the writer apparently wished to convey, so as to better

understand and evaluate any error. The need for this has long been recognised in error analysis research within Applied Linguistics (Pit Corder, 1975), and of course our teachers, in order to evaluate the essays, are in effect performing a kind of error analysis as part of that process. For instance, T1James says:

> *I'm not sure what she means by "high calories". I guess she means "high calorie foods".*

Other subtypes of this kind of interpretation focused not on words but concerned identifying features of the text as a prerequisite for then evaluating them. One of these is discerning rhetorical structure. For instance, T6 John said:

> *We've got an introduction, we have a conclusion, and we have the main body of the statement.*

Also T5Gena:

> *I think it's a new paragraph.*

Furthermore, there was a considerable amount of identification and classification of grammatical errors along with evaluation of them. E.g. T6 John:

> *Just carrying on, "In this year the pollution in seaside" – needs to be "at the seaside", again prepositional and article mistake.*

This was often accompanied by editing / correcting the error, as discussed in 4.7.3. Such corrections might or might not also be written on the script as feedback to the writer, but it was outside the scope of the study to record that.

It deserves comment here that, from the think aloud data, during stage 2 reading T3 Zain and T5 Gena produced more comments that are interpretation and not evaluative, compared with the other teachers.

4.8.3 Stage 3: Overall impression while summarising and reinterpreting judgments

This stage is the final stage of the rating process, where teachers construct, state and justify their final decision (talked about in part already in 4.5.3). Relevant examples were coded in the Self-monitoring - Judgment category as either: summarise, distinguish or tally judgements collectively; articulate or revise scoring; articulate general impression; or main cause for a given score.

A clear example of stage 3 is this from T6John:

> *So again, out of ten for content I would say, because of the fact she's got the structure – the introduction, the main body and the conclusion, and two good paragraphs – I would give that a seven out of ten. So a six out of ten for grammar and vocabulary, a seven out of ten for the content, which means a thirteen out of twenty. And again, at this level that's a good mark, I think.*

Here he shows clearly how he bases his final score, on his own personal scale of 0-20, on two sets of criteria equally weighted: content (including organisation), and language (specifically grammar and vocabulary). He then adds a final overall evaluative comment. This is a process not unlike that of using analytic marking scales in the testing of writing. No non-text based factors (4.6, 4.7) are mentioned.

T4 Sam in the following example illustrates his thought process in arriving at an overall view of a descriptive task in much more detail. In places, he relies on counting features in the text, not just rating alone, but again the argument is all text based:

> *14 adjectives in 80 words. We can work out exactly what mark to give him from those. Let's go with 14 adjectives and modifiers which is good. What else we got. He is consistent with his tenses. Yes, keep it…. Simple present. So that is 1,2,3,4, again 5, 6,7,8, 9 nine simple ... that's consistency good. 9 present simple. Good. He then got passive. Which is good, I am marking for that. He is also using present perfect. On a whole this is an accomplished paragraph*

> *and present perfect I have to mark well for this because he is managed to... has he got a topic sentence? Let's go back to basics. "my best friend is Ahmad". So, he got his name "I met him ..." yes, I am going to give him a mark for topic sentence. He got that correct. "he has ..."he summed up correctly. he is not revisited the conclusion correctly. So, I can't really give him that. Conclusion sentences. How about his body. How is he ... He has he made it coherent and he has kept the work flowing. He kept his essay flowing. He is actually pretty good. His coherence is good I am gonna give him very good marks. Coherent.*

In this instance, much is said which could equally be regarded as stage two talk, interpreting and judging, but for the fact that it is all presented as justification for a final rating on a personal mark scale.

At the other extreme, stage 3 could be given in words rather than marks, and quite briefly, along with some summary justification, as in this instance from T5Gena:

> *It's not bad. It's not great. I think the main problem, though, is that he has misunderstood the question and he's answered something else. He's tried, and I feel he's on the right track, but I don't think he's there yet.*

Another example from T5Gena illustrates how the overall rating was often not only accompanied by text-based support, but combined with non-text based justification (cf. 4.6), and also talk about implications (cf. 4.7). We break this excerpt into numbered subsequences for ease of exposition:

1. *OK, so it's quite a good essay.*
2. *I know this student and I know she is quite good. She's very good, actually. Her writing is consistently good, and this sort of confirms that.*
3. *However, yes, there are some mistakes. Nothing major,*
4. *and we'll look at those mistakes in the next class.*
5. *But as I said, there's nothing major there. I can understand what she's saying. On the whole, there's a good degree of accuracy – grammar and lexical accuracy – and cohesion.*

While one might have imagined that teachers would 'logically' first determine an overall rating based on the essay text evidence, then perhaps support it with non-text evidence, then move to any implications, real teachers move more erratically from one of these to the other. In this instance, Gena deals with the summary evaluation in three disconnected parts, 1, 3, 5, with increasing mention of text based support for it. In between she inserts a non-text based confirmation at 2, based on student past performance, and an implication at 3, for the teacher's future teaching.

Finally, T3 Zain illustrates an interesting phenomenon in the following example. Although the teachers were not asked to provide feedback to writers, many clearly did, as we saw in 4.7.3, and in this instance the teacher clearly sees his rating of essays as so intimately connected to this, rather than purely for assessment requirements, that his stage 3 thinking aloud initially lapses into being addressed directly to the writer, as if they were present. The *you* referred to is clearly the writer rather than the researcher. Indeed, the researcher observed the teacher write these comments on the script. About half way through, the teacher then switches to addressing the researcher:

> *That is a nice essay. I am very happy of range of ideas you offer and the amount of detail. A very good introduction. Your grammar is also good. One thing please, pay attention to your handwriting. Sometimes it is difficult to the reader to decipher your words. ... It happens how many times; one, two, three. Yes. Thank you.*
>
> *So, if I am marking this essay. In IELTS, that would be ... I would say 6. The main reason is there is no thesis statement. ...He did not recognise or mention the second opinion at all.*

It is particularly interesting that the teacher here does, in the end, clearly recognise that two quite different ratings are involved here. He provides feedback to the writer which picks out some positive and negative features of the essay, and, while not providing any

overall score or other rating, mainly uses words with positive meaning (e.g. *happy, good*) to provide an overall positive impression. The teacher's private rating, conveyed to the researcher, however, refers to quite different, negative, aspects of the essay, and provides a score.

4.9 Individual rater styles

We have already seen some evidence of rater differences, e.g. in their beliefs about the nature of good writing (4.2), their degree of use of the IELTS scale (4.3), frequencies of reference to different criteria (4.5), and sequences followed (4.8). However, in their detailed transcripts it was clear that their decision to focus more on certain features or kinds of comment could be seen as manifestations of their broader rating approaches or what some (e.g. Sakyi, 2000) would call their 'reading style', while they perform writing assessment, which was particularly apparent at stage 2. Since, as we have seen in the sections above, very much more is utilised by our raters in the assessment process than just what they read in the essays, we however prefer the term 'rating style' to 'reading style'. Such styles could also be referred to as patterns of what in some literature are termed 'marking practices' or 'decision making behaviours'. This is also similar to what Ferris (2006), in the realm of feedback, refers to as the teacher's 'philosophy'.

In this section, we therefore attempt to answer **RQ 8**, across all the aspects we have considered in the above sections.

RQ 8: What distinct rating styles can we detect being used by our teachers?

In their reading and rating after stage 1, all participants share the characteristic of reading the essay (which is hardly surprising) and, notably, referring to a wide range of criteria: as we saw in Table 4-4, there are not a great many criteria which were referred

to zero times by anyone. Beyond that, however, six distinct styles were observed from the six teachers' think aloud data, showing that, in detail, each of our teachers had their own unique style. This however is explicable given they were not all required to follow one system of criteria etc. for rating/scoring essays.

These styles are described below and reflect aspects like how exactly the essay was read, who for, and what criteria seemed to be foregrounded. In general, the styles are not a matter of one teacher doing something that the others never did, but doing much more prominently something that was also found more widely.

4.9.1 Reading for errors then reading for the criteria

T2 Bob's style, as we observed in the TA and he confirmed in his retrospective interview protocols, was to read the essay twice at stage 2, in each case discontinuously. In the first reading he focused on the individual features of the text that struck him as good or erroneous in order as they appeared. In the second reading, he read paragraph by paragraph considering a set of criteria in turn, leading to an overall rating.

The first reading is illustrated by the following excerpt. He read the entire text aloud pointing out the students' mistakes as well as correcting them throughout most of the process, sometimes on the script as well as orally. There were also certain occasions when he just commented on the error without correcting it, saying *it is not serious, small thing* or *odd*. Almost always, however, he corrected it.

> *"PAPER 1: There are wide range of applications" – missing word – "a wide range …benefit" – "which can benefit", no "in" – ".... people in the different ages". Not very good: "of different ages". Spelling of "different" is wrong. "…that younger people and old generation" – "the old generation" .... "For instance," – spelt wrongly – "students who study at university generally do research by computer" – "on the computer", a small thing….*

On the second reading, he progressed paragraph by paragraph and paid attention in turn to deriving overall judgments for key criteria such as paragraph structure, covering the required topic, vocabulary, grammar etc. (similar to IELTS criteria). This teacher's comments were focused entirely on features such as appropriate organization and vocabulary and the impression of frequency of errors such as repeated article mistakes and spelling mistakes. During his reading, he always reminded himself of the question and referred back to the task prompt. Here is an example of the second reading, with key words highlighted to show the criteria:

> *This is going to be difficult to assess. OK, good news is it's on topic and it's that bit has been done well, actually. I was a little bit worried about it in paragraph two, but it became clear in the end. Paragraph three is nice and clear. I'm going to give that an "as expected". The organisation into paragraphs is right; that's been correctly organised and we've got a conclusion. It's been finished as well. It's all doing what it should do. Linking? Actually, linking's pretty good. It's not over-used. There's a bit of variety there. How good is it, though? We've got those "for example" problems, which is not good. That's only one mistake, though, really – he doesn't know how to use "for example". What about other linking? What have we got? I like the fact that he hasn't overused linking. [...] I'm going to give it an "as expected". Vocabulary level and mistakes? ....*

We can see clearly how he goes over an agenda of criteria which he has in mind, considering how well the essay does on each in turn: topic, organisation, linking, vocabulary level and mistakes, grammar. Although the second reading is presented as focussed more on rating than the first (… *This is going to be difficult to assess*.), in fact the first reading is also full of evaluative comment about what is right or wrong. The difference is that the evaluation of the essay on the first read is driven and organised by the information source of the essay itself, while the evaluation on the second read is driven by the information source of the criteria.

In spite of his declared long experience in assessing writing, he seemed to read carefully and take his time in reading or rereading.

4.9.2 Reading twice with "three-category" focus

This is T6 John. In this style, there are again two readings, done in parts, but both are focused on criteria rather than just the sequence of errors in the text: the first, he himself states, is focused more on criteria of grammar and vocabulary, the second on the criterion of content. Hence it appears to involve three classes of criteria. In reality, the TA transcript shows, however, that the first read also covers spelling and the second also covers organisation, making five criteria in fact. This style is notable for a lower overall rate of mentions of specific criteria than the others.

He usually has a quick skim of the paper first at stage 1, where he counts the paragraphs, the lines and examines the layout of the paper. Then moves to the first read, focusing on his first two (or three) criteria:

> *So here I see I've got from Douar one piece of paper on two sides, written double spaced. I can see that she has written quite a lot, one two three… There are four paragraphs, it is handwritten. It is quite neat and readable. I'm just going to go through this now for grammar and vocabulary. …*

He then sums up his evaluation of those criteria

> *Bearing in mind this is B1A so it's lower intermediate level, it's not bad. It's the same sort of problems I'd expect to find when I mark writing: prepositions, plurals and articles. And because this lady is an Arabic reader and writer, some words are back to front, or some letters are as well. So, it's the sort of things I'd expect to find, and at this level that's actually quite good, because the mistakes are not that many. So, if I were to mark the grammar and vocabulary out of ten, I'd give that a six.*

The second reading he devotes to the content, but very much structured in terms of the main elements of the organisation – introduction etc.:

> *I'm now going to re-read this for content, and as I've described previously, we've got an introduction, we have a conclusion, and we have the main body of the statement. I'm happy with the introduction and the conclusion, because the way I teach them at this level is really just to only use one or two sentences for the introduction and the same for the conclusion, so that's OK. So, I'm just going to look at the main body now to see what sort of detail she's included. So here we have, just looking at the paper again, she starts off by saying, "In 1990 it was in the north", which is correct; she's interpreting that OK….*

Another feature of this style (in common with T1 James but not the others, for the most part) is that reports on feedback given to the writers (typically written on the script) are often made, rather than evaluation comments being all addressed direct to the researcher. E.g. during the first read:

> *Misspelling of "destroyed" – no "y" in there, so I've put that in. "Instead of these a factory, a supermarkt" – "ket", she's put "kt" – "and three residential buildings" – plurals are needed, she's put "residential building" – "have been built. Moreover" – which is a good use of the word "moreover"; I always put a tick there if it's a particularly good use of grammar or vocabulary…*

This approach locates the researcher as listening in not so much to the teacher's thought processes as to his communication with the writer.

4.9.3 Two full reads and focus on quantification of genre related features

T4 Sam is distinguished from other participants by reading through the whole text in advance, apparently continuously rather than in chunks, just to familiarise with the text before evaluating. Unfortunately, though, as we reported earlier, this was done before the TA session so the researcher was not there to record it and observe exactly how it

was done and whether it included any evaluation. In particular, we cannot tell for certain if that reading qualifies to be interpreted as just an 'initial impression' read, so part of stage 1.

In the TA session, there was then another reading, part by part. This read was done just following the essay and commenting on errors and good features of a wide range of types as they appeared. But unlike with the other participants, there was noticeable attention to counting up certain features of the essay: in particular, the word length, number of adjectives, and number of present simple tenses. They appeared to be chosen due to the fact that the assignment for his class (unlike the others which were argumentative) was a descriptive essay, in which adjectives and simple present forms might be expected. The attention to genre echoes what he said about features of good writing (Table 4-2).

This is illustrated in the following excerpts from the TA of one of the essays:

> *"He is very hard working and honest". Good two adjectives just what we need for our descriptive. Is this a modifier, this is the first one: 1, 2, 3. This is a modifier. 2, 3 we have to make a note of these. Is there any others before I go any further? No, ok.*

Later in same protocol:

> *"He has two clever children". Good. What is this more adjectives. Nine until now…*

Later again:

> *We can work out exactly what mark to give him from those. Let's go with 14 adjectives and modifiers which is good. What else we got. He is consistent with his tenses. Yes keep it…. Simple present. So that is 1,2,3,4, again 5 ,6,7,8, 9 nine simple .. that's consistency Good. 9 Present simple. Good*

This procedure is clearly also related to how he arrives at an overall score, which is based on a numerical calculation rather than an intuitive summary of a lot of specific ratings. Thus, while other teachers might refer to errors or other features being frequent or repeated, only Sam actually counted them up.

4.9.4 One detailed read, focus on feedback

T1 James in contrast to the above really does one reading only, with little rereading. He reads paragraph by paragraph, introduction, body, conclusion correcting errors at all levels, with IELTS criteria in mind, as he is the IELTS trained participant: task achievement, coherence and cohesion, vocabulary and grammar. He gives the most detailed account in terms of numbers of mentions of various criteria.

It is noticeable also that his think aloud largely takes the form not of talking direct to the researcher about what he is thinking, but rather reporting to the researcher the feedback to writer which he is usually also writing on the script, as John also did to a lesser extent. This feedback consists sometimes of direct corrections, sometimes use of codes like WF for word form, and also overall comments addressed to the student at the end. While the other participants do sometimes indicate what feedback, they are giving or would give to the writer, and indeed T3 Zain wrote a lot of feedback on the scripts, for them that is presented in the way they talk as secondary to the evaluation being arrived at by the rater for him/herself, and for the researcher. This is all the more remarkable since he is the only participant with IELTS examiner training, and of course that training is very much focussed on the role of testing rather than teaching. We can see this for example here where he extensively reports what feedback he is giving rather than addressing the researcher more directly:

> *I'm just correcting "as": it should be "has become an epidemic and it constitutes a real threat". Also, there's a spelling mistake: he's spelt "health" without an "a", so I'm just going to write "SP" there. OK, then he says, "It seems that the general public does not realise that the real dangers." "Does not realise that the real dangers" doesn't make sense, so perhaps "does not realise the dangers", so I'm just crossing out "that" and "real". This is OK. I think, however, that he needed to mention something about the question, really, so perhaps to say, "There are various reasons for this; however, I do believe it can be solved." So I think he doesn't need to answer the question, but he needs to refer to it at least here, so I'm going to add that. So yes, I'm just going to write, "There are a number of reasons for it, but I believe it can be combated." I'm using the word "combated" because this is a word that I often teach to my IELTS students, because it gets very boring when everybody writes "solve, solve, solve" in terms of problems.*

In this approach, the teacher reads, apparently holding in his mind well-defined IELTS criteria in his head. They are not often overtly mentioned but clearly lead to his IELTS score given at the end. Mostly, he gives immediate judgments as he reads through and at the end he gives his final evaluation in a leaner way by focusing on four criteria with all compositions.

4.9.5 Much reading and rereading, with varied focus depending upon teacher's expectation and the essay itself

T3 Zain basically performs one read, in many chunks, with quite a lot of rereading of chunks. Indeed, he does way more reading and rereading of parts of the essays than any other teacher.

A range of criteria are usually used in evaluating any essay, primarily language focused, (cf. 4.5.1) but the focus of attention differs from essay to essay, depending on the essay and, to an extent, on expectation based on prior knowledge of writers. As T3 Zain says after reading the first essay:

> *Well the first student was so grammatical accurate actually. I focused mainly on ideas and structures of paragraphs. So, I expected that. In, om om other essays, that student was kind of exceptional. So, other essays, it will be more comments on grammar,*

Again, later he admits the potential for bias, but claims to adjust his expectation based on the writer, or type of writer, by use of the information source of the actual text itself:

> *I might have preconception. So, if I am marking you know well from my experience with teaching Saudis, they have got spellings and grammar mistakes but there are exceptions. So even you have this conception in mind sometime you change it when you first read the first paragraph.*

As a consequence, although criteria from all levels are used in most essays at least some of the time, there are declared focuses of attention: essays 2 and 5 grammar; 1 and 6 content; 7 handwriting; 3 and 4 no clear specific focus.

4.9.6 One read, with high level criteria focus

This pattern is evidenced by T5 Gena. At stage 2 T5 Gena generally read the essays once, not as a whole but part by part, without much rereading of parts; only on one did she say *I would need to read it again*. As she did so she made references to good and bad features relating to criteria at all levels from spelling through to content, but it was clear that what was uppermost in her mind all the way through was the higher levels. She repeatedly foregrounded the organisation /rhetorical structure and aspects of the content, thus giving the feel of a top down approach, in terms of levels of language.

First, units of rhetorical structure were constantly referred to as the basis for chunking the text into units to be read and rated at one time. This was also followed by several of the other teachers but is much more prominent in T5 Gena who used the words *introduction* and *conclusion* far more frequently:

> *Then he goes on, and I can see he's divided the rest of the essay into two more paragraphs and then a conclusion, so he's followed the typical structure, organisation that we would expect from an IELTS task 2 essay.*

Furthermore, in many instances attention is paid first to the content of the element of structure being considered, rather than the language, here going into the body:

> *OK, he's starting off his second paragraph by saying he's going to explain these reasons, or there are several reasons. "Firstly, students can read a lot of articles through the internet wherever they may be." That's good.*

Again, the argument is prioritised here:

> *Yes, he's making a good point, but he's not giving – I don't know, he's not really developing it. His language is not bad.*

Content is addressed not only in terms of argumentation but also in terms of task fulfilment:

> *So, the next one. Right, I can see straight away there's a problem here. The student begins, "It is widely believed the internet does not able students nowadays being better than students in the past. Personally, I completely agree with the students in the past are better than the students in the present." Now, I'm just looking back at the question to make sure I've – "the internet enables students today to be better informed than the students of fifty years ago" – and he says "It is widely believed that the internet does not able students nowadays". So he's misunderstood the question. That was what I was afraid of. He obviously hasn't understood the word "enables",*

Errors of grammar, vocab, spelling etc. are looked at extensively but often viewed as subordinate to communication success. In this excerpt, for example, a negative language point is immediately softened by reference to its communicative success:

> *"On a negative side" – OK, that's something wrong there. Yes, I think what he means there is "in a negative way", but I understand what he's saying.*

Again, in this instance the higher level is weighted above the lower:

OK, there's a few things, mistakes, but overall that's another good paragraph. He seems to have good control of the language and he's expressing his ideas; he's giving examples, which are interesting examples and they make it, the whole essay, they make it interesting.

4.10 Effect of teacher background

We finally briefly address the last **RQ**:

RQ9: Is there any difference in any of the above between individual teachers, according to their general training, experience of rating writing, prior knowledge of a specific rating system, etc

The account above has revealed considerable variation between teacher-raters, particularly in relation to use of criteria (4.5), and overall process style (4.9). Indeed, while all teachers engaged in the same kinds of reference to criteria, non-text based support, other comments, and three stage process, in important ways, each teacher was different. We have already mentioned in the account above any instances where teacher background seemed to explain such teacher variation, so we will simply refer to the main points here again in brief.

Overall, perhaps the main finding is that much of the variation we observed was not obviously explicable from any of the background characteristics of the teachers that we knew of. Hence much of the variation might be regarded as a matter purely of idiosyncratic individual differences, or perhaps of background features of the teachers which our study did not discover. Really just two main background features did seem to have an impact: these were the IELTS examiner training of James and the non-native teacher status of T3 Zain.

The fact that T1 James had had IELTS examiner training clearly affected his behaviour in that, of all the teachers, he adhered far more closely to following IELTS criteria and an IELTS procedure of analytic rating, and used the IELTS scale routinely for reporting. The other teachers often only gave an IELTS score when the researcher expressly asked them to do so in the retrospective interviews. That training also probably showed itself in his ability to apply criteria to more specific instances in the student essays, or at least to verbalise that, more than the other teachers. Nevertheless, this training did not show itself in his treating classroom rating as just another case of IELTS exam rating, since of all the teachers T1 James was perhaps the most inclined to see the activity as having the purpose of giving detailed feedback to the writers, which of course cannot happen in a real IELTS exam.

T3 Zain was the only non-native English speaking teacher included and this seems to have impacted on his behavior in several ways. It possibly accounts for his doing more reading and rereading of parts of the texts than any other teacher, since of course he had greater difficulty in some instances with understanding what they had written: he evidenced his own language limitations with respect to the word *fruition*, for example. Furthermore, his religious/cultural background as a Syrian Muslim also showed itself in his personal responses to the content of what some of the writers wrote in their essays. We might also argue that the great amount of detail in his response was largely influenced by his desire to give extensive feedback to the students, which in turn may be due to the fact that, as a non native speaker himself, he had had to take IELTS writing tests in the past and so had a more immediate idea of what would help the students.

4.11 Chapter conclusion

This chapter has provided answers to all the research questions, which will be summarised next in the Conclusion chapter. Overall the picture of the rating process which emerges exhibits more variation between raters than is found in the more usual kind of study in this area which approximates more test or exam conditions. Furthermore, it reveals a much richer range of mental activity being reported, particularly in the area of matters that are not founded solely in the essay text product and not directly related to evaluation. Nevertheless, the range of types of criteria reported and the basic three stage model of the rating process do match such other studies.

In conclusion, it is apparent that the account above in this chapter, since it proceeds RQ by RQ, necessarily disperses the information we obtained about each individual teacher. Hence, we feel it illuminating to revisit each teacher and present a case by case account. In this way, unfettered by the focus of a particular RQ, each teacher's findings can be dealt with in terms of their own narrative and the thoughts and reasoning behind their reading and rating behaviour can be dealt with in their own individual terms.

4.11.1 Bob

Bob is a highly experienced male NS TEFL teacher in his 50s, with BA, and some training and experience in grading Cambridge exams at C1 and B2 levels. He believes his intuition based on experience is a better guide than training, however, enabling him "*to recognise certain patterns, and therefore .... able to assess things pretty quickly*". Therefore, he sees rater development as occurring not through training but mainly

through teacher experience in a particular context where the teacher teaches, leading to a rating procedure that might be unique to that situation.

Consistent with that belief, he does not simply adopt a standard rating scale but is eclectic. He refers to IELTS criteria (since he is teaching an IELTS class) and to his perception of what is CEFR B2 performance (since that is the assigned level of the class), but adds his own special attention to genre as a criterion, and his own weighting of relevant ideas, and organisation, above achieving a word limit: "*First I always look if the student answer the question, and the organization. Grammatical range is less important for me than grammatical accuracy and organization. I think if I have those I think then I can develop the others".* He also, more idiosyncratically, prefers handwritten to typed scripts because he believes *"psychologically, if it's printed, it gives the impression that it's going to be right somehow; whereas if it's handwritten, it psychologically inclines you to think it needs more correction."*

As far as his practices go, Stage 1 is often short, but always involves recalling the writing prompt. It might be followed by activating what he knew about the previous performance of the writer: "*Now, Hakim, I don't expect this to be all that good.*" Also, there may be a comment on overall length, e.g. *"It looks like he's done quite a lot".*

In stage 2 he reads the text in parts, not as a whole, with a good deal of rereading. He goes through the whole text twice.

The first time, he reads fully, pointing out the students' mistakes as well as usually correcting them, sometimes on the script (so creating feedback) as well as orally (for the researcher), in order as they appeared. He may also comment *it is not serious, small thing* or *odd*; *basic grammar mistake, lack of care of building blocks, completely off topic.* He considers the effect on the student. Overall, he is highly pedagogically aware,

looking for reasons for errors, such as L1, writing like speaking, past classroom engagementof the student.

In the second reading, he uses the text to support his judgment of it on each of a set of criteria which he has in mind, considered in turn: coverage of the required topic, paragraph structure/organisation/linking, vocabulary, grammar, spelling (based on perceived general frequency of error) similar to IELTS criteria. He was most frequently coded for reference to topic relevance, paragraphing, cohesion/ coherence, and vocab and spelling errors.

At stage 3 he is reluctant to combine his verbal (not numerical) assessment of the essay on each criterion into a single overall score (doing so only when pressed by the researcher). Instead he prefers to encourage the students to meet their target, e.g. *need to work harder.* He works on ratings to serve both himself and the student. Overall it is clear that decisions on feedback mostly are made after those made in the process of arriving at ratings for the writing itself, and are kept separate. After reading at stage 3 he summarily updates his knowledge about the student: *Better than I might have expected from him to be honest, Amjad.*

Overall his account is shot through with pedagogical references to things he might teach, what he has taught and so the students should observe, what he expects of the student, how he is updating his view of the student, etc. This contextual embeddedness of his rating makes it very teacherly rather than like a tester, and no doubt reflects his long experience.

4.11.2 Gena

Gena is a British female in her 50s. She has a BA in History and DELTA training, together with more than 20 years of TEFL experience. Like Bob she does not highly value training, especially if it is repetitive and not introducing anything new, like the departmental training sessions she mentions attending. Rather, she believes it is better to rely on 'common sense' and 'instinct', which of course implies using what she has learnt through experience and reflection.

Consistent with that, when she talks about her criteria, she says she does not just adopt a standard rating scale, such as IELTS (even though she is teaching an IELTS preparation class). She reports valuing some features of writing which are not standardly listed in marking rubrics for writing, such as lack of errors due to negative transfer from L1, quality of presentation in the form of handwriting and layout, and the writer engaging in a planning phase before composing. Something these share is that they all impact directly on how well the text can be understood by a reader, which turns out to be her emphasis when rating the scripts in practice as well as her reported belief.

She quite explicitly recognises her purpose in these class writing tasks as being "*to help students to write better more than to put them in exam conditions where they expect a score."* I.e. she sees herself as a teacher more than an examiner. She shows awareness also of the communicative purpose of the writing they do: "*I'm reading this as a teacher: I'm marking it, evaluating it. But somebody who's just reading it, if it were an article in a newspaper or something, ... you can't focus on the message because the mistakes get in the way. ....".*

Her actual practice follows the usual three stages, though implemented somewhat less thoroughly than by Bob. At her first stage, she reads the task prompt, though not for every script, and makes some predictive comments based on scanning the name and the

script. E.g. *"Right, the third one. OK. This one the handwriting, it's very neat but it's always a little bit difficult to read her handwriting. So again, I know who the student is."*

Stage two consists of just one read, in multiple parts, which are usually units of rhetorical structure. She does not comment on every error as Bob did, nor give the impression that her TA comments to the researcher will also constitute feedback to the writer. The main coded areas of comment are relevance of content, introduction, and other aspects coded in the general rhetorical and ideational categories: "*OK, he's starting off his second paragraph by saying he's going to explain these reasons, ... That's good."*

There is some attention paid to vocab choice, and error frequency, but overall less attention paid to language errors compared with the other teachers apart from Sam, and greater use of the words *introduction* and *conclusion*. Errors of grammar, vocab, spelling etc. are often viewed as subordinate to communication success. Gena also refers to her past teaching/feedback, often showing additionally some irritation: "*I keep telling them".*

Stage 3 is primarily given in words rather than marks, and quite briefly, along with some summary justification which again shows that her agenda of criteria is led by the higher levels rather than language details e.g. "*He answers the question – he misses something, as I said, but his argumentation is good."* She does award scores, although, as she says "*IELTS is different, the scoring. I don't do IELTS marking.*" Her scale (marks out of 10) is partly modelled on IELTS, for which she admits she is not a trained examiner, but claims instinctive familiarity through experience. She also takes into account a scale provided by the organizer of the course which she teaches on (a scale

which in fact uses criteria quite similar to IELTS), together with her own added criteria. The marks seem to be for her own benefit rather than to be communicated to the students.

Like Bob, Gena refers to her knowledge of the writer both before reading and after reading, where she updates her knowledge of the student but typically allows the evidence of the script to have the last say. Sometimes however, in trying to award her summary rating she seems to be influenced by what the writer needs rather than what the essay deserves: *"I'm not sure if that would be enough for him if he wants to go on to study in a university. He's still got some work to do. It's a difficult one."* Also in teacherly fashion, Gena expresses her views about the achievement of a student in a way that presents him not as being deficient but rather as in the process of learning - on a developmental trajectory.

4.11.3 James

James is in his 40s with 18 years of experience. He is TEFL trained and is an IELTS examiner. He retrains for this every two years, with detailed attention to each IELTS criterion, and this clearly has a big impact on his entire rating behavior in our study.

As he says, and does, "*Because it is an IELTS class, I am using IELTS criteria*", and the only additional criterion he mentions is the composition 'being interesting'. Unlike Bob, he values top down training rather than just relying on experience, and unlike Gena, taking the stance of an examiner rather than a marker of practice assignments, he believes the repetition of the training to be valuable, since it enhances inter-rater agreement. He also thinks the training increases his confidence and expertise, and the researcher felt it possibly helped his TA fluency as well. In his practices, however,

while he does stay close to IELTS rating, he adds many teacherly aspects to the rating process.

In stage 1, he may read the prompt and often recognises the students and activates his background knowledge, as the teachers generally do. "*Who is that? This is Barbra from Turkey. Umm, as far as I know she got problem in grammar I think."*

In stage 2 he reads once in considerable detail, part by part, with some rereading. He clearly is bearing in mind IELTS criteria throughout. The depth of attention paid (or at least the degree to which he reports it) is attested by the fact that his reports contain far and away more references to various rhetorical/ideational and language features than those of any of the other teachers. In the former category, paragraphing and cohesion/coherence are most often targeted, in the latter, vocabulary, spelling and general lack of clarity.

Notably he frequently refers to what he is giving as feedback to the student as he speaks to the researcher. This feedback consists sometimes of direct corrections, sometimes use of codes like WF for word form, and also overall comments addressed to the student at the end. He always maintains a clear distinction between the analysis and judgment made for himself and the researcher, as against what he will communicate to the writer.

At stage three he sums up the relevant criteria and produces an IELTS score, though not through the strict procedure of rating each subscale separately and combining those ratings. This is not part of the feedback to the writer, however: "*I am not writing scores on their paper*". For the writer, he often voices what he would say as qualitative summary feedback: "*So, the biggest problem Hakim has here is that there's a lot of irrelevant stuff, ... So, I'm going to write, "Not bad writing, Hakim, but you have forgotten about the question in two paragraphs..."*

Occasionally in stage 2 but more thoroughly at stage 3 he refers to aspects such as the L1, own level, classmates' level, and recent performance of the student. He further uses such contextual information to give the scores he awards more meaning for himself: "*So, bearing in mind he's only a B1 level student…. so, to get a 5.5 or a 6 is a very good score for one of those guys. This is a very good essay for somebody in this class....".*

Thus, in spite of his IELTS examiner background, James allows his instincts as a teacher to come in. Indeed, he shows more awareness than other teachers in commenting explicitly on this: "*.... I care about my students, and so I'm pleased to see, "Oh, look, Ali did that....*". And referring to feedback that he would give: "*Rather than saying "He's done this and this and this wrong", you could say "Well, he's only done this and this and this wrong". He hasn't made as many mistakes as many of his classmates might have made."*

4.11.4 John

John is a British male in his 40s with a PGCE and TESOL certificate, and 15 years of TEFL experience, but no special training in using professional rating scales. It is apparent, however, that his general training did cover error analysis, which is relevant to writing assessment. He claims balanced benefit from both training and experience: "*Training gave me the chance to reflect on my experience in writing assessment in a more positive way..... But of course, the other thing is, from an ongoing point of view, I mark every week so it's an ongoing learning process.*"

He consequently claims to use only his own personal criteria, which in fact closely matched the list offered by the researcher apart from special attention to genre. Indeed he argues against adopting any fixed scheme: "*Fixed criteria can be more reliable than*

our own ones and sometimes it is fairer to use them. But knowing the students' needs and level, using a fixed scale can lower their grades from the overall grades they deserve. ... Sometimes, I prefer a mixture of criteria, where I can be present."

This statement is consistent with his awareness from the start of the difference between assessing these practice essays and exam essays: "*In this class, we are not testing students' proficiency but we are trying to develop their writing skill in a more pleasant way,..*" He is in fact not taking an IELTS exam preparation class but a general intermediate class. This colour his whole process.

At stage 1 he sometimes reads the prompt, then usually has a quick skim of the paper and may count the paragraphs, the pages, and examine the handwriting and layout. Like other teachers he calls mind some background information about the writer: e.g. "*This is Doura this lady is an Arabic reader and writer, some words are back to front, or some letters are as well. So, it's the sort of things I'd expect to find. Let's see.*"

Stage 2 is characterised by reading twice with 'three-category focus', each time part by part. Hence, he is also coded for quite a lot of rereading.

Both readings are focused on criteria rather than just the sequence of errors in the text: the first, as he overtly states, is focused more on criteria of grammar and vocabulary (together with, in practice, spelling), the second on the criterion of content (including in practice also organisation). He records the lowest number of coded mentions of specific criteria, though the most frequent are content relevance and spelling.

He anticipates stage 3 by in fact summing up the first read at stage 2 before moving to the second read: "*Bearing in mind this is B1A so it's lower intermediate level, it's not bad. ... actually, quite good, because the mistakes are not that many. So, if I were to mark the grammar and vocabulary out of ten, I'd give that a six.*"

The second reading he devotes to the content, but very much structured in terms of the main elements of the organisation – introduction etc. Like James, he reports on feedback given to the writers (typically written on the script, and including positive ticks) as much as addressing evaluation direct to the researcher.

At stage 3 he is well able to externalise his process of arriving at an overall judgment, in practice weighted in favour of the first read criteria: "*So I'd give that one a 7 out of 10, mainly because there aren't that many mistakes, for grammar and vocabulary.*" He also gives weight to general comprehensibility.

Even more than the other teachers he introduces at stage 2 and 3 many teacherly remarks making connections with, for example, the writer's level and previous performance, and essays written at the same time by different writers of the same level. He also connects with his past and current instruction, and indeed how what he learnt from the current essay rating will guide his teaching of the student in the future.

4.11.5 Sam

Sam is an English male in his 40s with CELTA, an MA in Linguistics, and 15 years of TEFL experience but no special training related to writing assessment. He resembles John in his background, and similarly is teaching a general intermediate class rather than an IELTS class. Consequently, he too claims to be using entirely his own criteria and scale (in his case with 33 points), which he sees as a strength: *"When you've worked with students, because this is not a credit-bearing module, I can make my own criteria and I can be a bit more generous ... it can motivate the students".*

He believes in special attention to genre appropriacy: "*If it is a descriptive text, the student needs to use descriptive language and include adjectives*"; good writing "*should*

follow the given rubric and be planned according to it." His practices reflect his focus on genre, although another belief stated at the start, that (like John again) he is "*keen on grammatical range and lexis*" is not reflected in practice with respect to grammar. He does however recognise his variability "*If I'm honest, the one where they can score the most points in the hierarchy is vocabulary.*"

At stage 1 Sam differs from the other teachers in first reading through the whole text apparently continuously rather than in chunks, just to familiarise with the text before evaluating. This was unfortunately not observed by the researcher but there is evidence later that he probably also read the prompt and recalled the background of the writer at this stage.

At Stage 2 he reads in detail, dividing each script into only 3 or 4 parts to read, compared with other teachers typically 6. This is related to his choice to set a descriptive paragraph task, not an argumentative essay in the style of IELTS, which the other teachers set. Hence Sam's scripts were shorter (around 80 words rather than 200).

He reads following the essay and commenting on errors and good features of a wide range of types as they appear. He is coded most often for mentioning vocabulary choice, written quantity, and text development. He additionally counts up certain features of the essay, especially word length, number of adjectives, and number of present simple tenses, which he deems especially relevant to the descriptive genre of the assignment "*Good two adjectives just what we need for our descriptive.... Is this a modifier, this is the first one: 1, 2, 3. This is a modifier. 2, 3 we have to make a note of these...*" He also mentions a criterion not highlighted in advance as a belief: consistency of use of present simple tense.

At stage 3 it is clear that the counting done earlier is related to how he arrives at an overall score, using some numerical calculation rather than a purely intuitive summary of performance on separate criteria: *14 adjectives in 80 words. We can work out exactly what mark to give him from those.*" However, in detail he exhibits his own weighting of criteria, for example giving no weight to mechanics for the current writing task, although he did not indicate that in the interview. Furthermore, he operates a flexible policy with respect to the influence of different criteria, e.g.: "*I have decided to give her extra couple of marks because she stuck to the word count* " and "*he's made three paragraphs and there should be one. .... For that reason, I can't give him a good mark*". He also refers to a later possible final revision of marks: "*I am going to revisit this .... when I read all the scripts*."

In addition, he evidences a number of teacherly practices such as reference to where a student seems to have paid attention to recent teaching, and comparison with prior performance of the writer. Finally, he is the teacher who refers most to the affective dimention of rating writing both for the student (reference to motivation) and himself: "*He's using these more complex noun phrases to describe his friend rather than just single or paired adjectives. It's really encouraging actually...Because it makes me feel good as a teacher.*"

4.11.6 Zain

Zain differs considerably from the other teachers in being younger, in his 30s, and a Syrian native speaker of Arabic with a PhD. He has some teacher training, broadly including assessment, which he feels is useful alongside his experience: "*My focus during writing correction as well as my understanding of some aspects of writing features, have developed after training*."

Since he is teaching an IELTS class he claims to use IELTS criteria as well as some of his own: he is unique in explicitly referring to the communicative criterion of audience awareness: "*the student should know what the teacher expects them to write. I always tell my students to put themselves in the place of the reader, what do you want to read, what knowledge should you share?*"

Like Sam and others, he holds the belief that criteria should vary with genre: "*We have to be careful in choosing suitable criteria because each genre has its own criteria, although there are some common features among different genres".* He also thinks they should vary with student level.

At Stage 1, like most of the teachers, he reads the task prompt and activates his knowledge of the writer to guide his rating: "*Well this is Dali, she writes some complex sentences in her previous compositions. So, let's see what she produces.*" He also uniquely refers to using essays already assessed to guide future rating: "*the first student was so grammatically accurate actually. I focused mainly on ideas and structures of paragraphs. ....., that student was kind of exceptional. So, other essays, it will be more comments on grammar.*"

At stage 2 Zain performs one read, but with twice as much reading and rereading as anyone else, reading shorter chunks. This seems at least partly due to him being a non-native speaker of English so finds it harder to understand the writing (there were a few words he had to look up) and harder to retain longer stretches in mind while evaluating them. Thus, many of his comments are coded as interpretation and not evaluative, compared with the other teachers.

He records the second largest number of coded references to criteria, though well below James. The most mentions are of relevance, introduction and reasoning among the

rhetorical/ideational codes and punctuation and general clarity among the language codes. The language references outnumber the others, possibly influenced by his experience as a foreign learner himself in Syria where teachers may focus on language use in writing assessment. As he admits, the precise focus of attention to criteria differs from essay to essay, depending on the essay and, to an extent, on expectation based on prior knowledge of the writer.

Another feature not prominent in the feedback from the other teachers was genuinely communicative response to the content of what was written. For example, in a couple of essays he disagrees with what the writer says from a religious/cultutal standpoint. Furthermore, though not asked to provide feedback to writers, like some of the other teachers he explicitly refers to feedback, but unlike them, goes to the extent of sometimes allowing his TA to lapse into addressing the writer rather than the researcher.

At stage 3 he arrives at an overall evaluation in words and a score for himself and the researcher but not the student: "*I am not giving them marks actually*". He further filters the evaluation in words that he will include in student feedback on the principle of how peers might react: "*I will try to avoid words like good and very good because they are sensitive if they see their colleagues got 'very good' and 'good'.*" Like Sam, he weights criteria in an ad hoc way depending on the essay and writer: e.g. "*Because it is grammatically good it does not matter if you write less than 250 words*."

Zain engages even more than the other teachers in teacherly comparison of performance with what student did before: "*I keep saying this student should not be in C1B which is the advanced level. I would put him in intermediate or upper intermediate*." He further

updates his own knowledge about the student: "*I am surprised because one of the teachers mentioned that he is bad in listening and speaking.*"

We might conclude that the extent of detail in his response is influenced by his desire to give the richest possible help to the students, which in turn may be due to the fact that, as a non-native speaker himself, he had experience of taking IELTS writing tests in the past and a greater empathy with them generally. For him, that outweighs any use of the assignments for assessment purposes.

4.11.7 Overall conclusions concerning the six teachers

Summing up the perspectives across our participants and our research questions, we may make the following generalisations. For the most part, what the teachers said about what they did in the general interviews, reflecting their beliefs, did accord with what emerged from the TA and follow up interview sessions, which we regard as closer to reflecting their actual practices.

Our teachers largely agreed on the core features of what they considered to represent good writing (broadly task completion, organisation and cohesion, grammar, vocabulary, and mechanics) but they mostly claimed to, and actually did, each add some criteria of their own, especially attention to genre differences. They also all broadly followed a three stages process of first scanning the composition and activating their background knowledge of the writer, then reading and applying criteria analytically, before finally summing up and combining ratings of separate criteria into an overall assessment which in most cases they did not wish to transmit to the writer as a score. Most notably during stages two and three they all engaged in varying degrees in comment which went beyond assessment into the realm of teacherly concerns including

relating the performance to that of the student and peers on other occasions, providing feedback for the writer, and suggesting implications for teaching. The similarities are notable given that the teachers differed widely in relevant training, and in their views of the value of training in comparison with experience.

The main differences were to be seen in the precise styles of reading which they employed at stage 2 (number, focus and detail of readings), the weightings chosen for different criteria in arriving at an overall rating at stage 3, and the scales used to record summary assessment. This variation could only partly be explained by differences in teacher background in terms of training and NS/NNS status.

CHAPTER 5 CONCLUSION

5.1 Introduction

In this chapter, we will first summarise the findings which we obtained, and their interpretation, with respect to the research questions. Next, we will point out the significance of the study and suggest implications of the findings for participants in the situation where classroom writing is assessed. After that we recognise the limitations of the study, and make suggestions for further research.

5.2 Summary of findings

Here we briefly revisit each research question, indicating what we found.

RQ1: What do our English writing teachers perceive as good writing?

In answer to (RQ 1) we found that the teachers did not vary much in the core features that they claimed to recognise as typical of 'good writing'. Cohesion, task completion, development of ideas, vocabulary, grammar and organisation all were strongly endorsed by all. This could be influenced by the fact that this was the only data gathered in the form of closed rating responses to the items offered, rather than open data. There was some variation on the role of mechanics and additional features suggested such as content interest, genre appropriacy, and planning. What criteria the teachers actually paid attention to when rating in practice varied much more than was suggested by the answer to RQ1 which perhaps reflected more their beliefs. This attests to the value of using multiple sources of data.

RQ2: What rating scale and criteria do the teachers claim to typically use when rating student writing, and what are their sources?

Teachers differed a great deal in the scales they reported using to sum up the rating of features of the essays or the essays as a whole. Scales ranged from something very close to the official IELTS scale and using full IELTS criteria to just using evaluative adjectives such as 'excellent' and personally chosen criteria, though often with an eye to IELTS criteria since several of the teachers were teaching IELTS classes. Correspondingly, while most claimed some effect of the IELTS scale and criteria on the way in which they rated class essays, all but one also admitted to using their own scale and criteria at least in part. Key themes which emerged from what they said were that they were mostly more concerned with the consequential pedagogical value of their rating for the students than with providing a test-like reliable score, consistent with their approximation to analytic rather than holistic rating practices, and that their criteria varied depending on the essay genre and even the individual student.

RQ3: What training have the teachers received that is relevant to assessing writing and what are their views on such training?

The sample of teachers evidenced a wide range of training, from IELTS examiner training to training with no special attention to writing assessment at all. Attitudes to training were mostly positive, indicating its value not only for awareness of criteria but also the process of rating, and teacher confidence. There were however differing views about the need for repetition of training, and the role of experience. These were interpreted as associated with the difference between examination rating and classroom rating: the former relies on top own uniform training in order to achieve high reliability between raters, while the latter is more concerned with being relevant to the students so relies more on teacher experience, bottom up.

RQ4: What are the most important qualities that the teachers look at in practice when they are rating their students' samples? (criteria used and their weighting).

Overall, two thirds of the 24 separate criteria coded were used by all the teachers but there were differences between teachers for instance in preference for higher level criteria (related to content and organisation) versus lower level (language related) ones. Overall, the language criteria of vocabulary and spelling, were most mentioned, closely followed by clarity, reasoning, argument, and quality of the introduction. All the teachers exhibited a kind of analytic rating, often referenced to other students rather than criteria defined in absolute terms. They also engaged in decision making behaviour where they weighed different criteria against each other in deciding on an overall rating for an essay (whether expressed as a score or not).

RQ 5: How do the teachers support, explain or justify their rating/scoring?

A wide range of information sources was used to support and justify the rating decisions made, beyond the essay text. They included the writing task description, background information about the writer and his/her classmates, what the teacher had taught the students, and what the teacher thought the impact of the rating might be on the writer, if they received it in feedback. These considerations were evoked not only after reading the essay but often before, in which case they coloured the whole evaluation that followed.

RQ6: What other kinds of comments do the teachers make, beyond those directly related to achieving the rating itself?

Four kinds of wider aspect were identified in teacher comments, which arguably were not central to rating or evaluating an essay, but rather invested the bare ratings with more meaning for teachers or writers. One was reasons offered for student performance,

which were usually related to prior knowledge of the students. Another was implications for the teacher and teaching, where the rating updated the teacher's knowledge about the writer, or suggested future instruction. The third was implications for the writer in the form of feedback, which in some cases seemed to be separated from the basic rating in the mind of the teacher, at other times not, and was often affected by how the teacher thought the feedback would be received by the writer. Finally, there were teacher communicative responses to the message conveyed, rather than to the language, task completion etc.

RQ7: What common sequence of activity can we find in the rating behaviour of our teachers? E.g. Always read whole text first? Consider criteria in a certain order?

Similar to Cumming et al., (2002) three broad stages of the rating process were identified. The first consisted of reading information such as the writer's name and the task prompt, and deriving information from background knowledge about the former, as well as perhaps skimming or just glancing at the text to gain an initial impression. The second consisted usually of reading and rereading parts of the essay, associated with interpretation and judgment of what was being read after reading each part, mostly text-based. The third was where a summary judgment was obtained, based on the essay text evidence, and wider non-text based justifications and implications might be alluded to.

RQ8: What distinct rating styles can we detect being used by our teachers?

The data showed that really every teacher had a different style, when variables such as how many times they read the essay at stage 2, what criteria they focused on, and how they talked about the criteria were taken into account.

RQ9: Is there any difference in any of the above between individual teachers, according to their general training, experience of rating writing, prior knowledge of a specific rating system?

Although the teachers exhibited some features in common, such as referring to much the same general range of criteria and following a three stage process of rating, there was considerable detailed variation in frequency of use of criteria and overall style. This however for the most part could not be simply explained from their background. Just James' focus on IELTS was explicable from his training and some of Zain's behaviour from his being the only teacher who was not an English native speaker. It is possible that some of the variation observed was more due to the differences in assignment topics and genres which were set by teachers and to the differing proficiency levels of the students. For example, there were differences between Sam's criteria and those of other teachers which seem to be explained by the fact that he alone set a descriptive writing task for his lower proficiency students.

5.3 Significance and implications

At the start of this study (chapter 1) I was at pains to make clear the distinction between the present study and research in contiguous but distinct fields. I believe the findings of the study have fully justified my premise that there exists an under-researched area of teacher cognition and practices distinct both from research into how testers (including teachers) apply set scoring criteria and scales to writing done in exam conditions, and from research into teacher feedback to students on their practice compositions. The significance of the study therefore has two sides, in that it has fundamentally

illuminated just how the kind of assessment that it focused on really does differ from that studied in neighbouring areas, and what it contributes.

5.3.1 Significance for testing research

First, the findings contrasted considerably with those of writing assessment studies (e.g., Vaughan, 1991; Wolf, 1997; Sakyi, 2000; Erdosy, 2002) of the more usual type, where the same essays are marked by all raters, using the same marking scheme (criteria and scale), in test-like fashion. For the research community, the contribution that we claim therefore is to have illuminated this in our study better than perhaps any other study hitherto. There are many specific differences.

One key difference from exam rating is that, in the kind of rating which we studied, where no rating scheme was imposed, teachers did not feel, either as reported in their beliefs or practices, obliged to simply adopt some standard rating scheme that they were familiar with and use it. This is remarkable since some of them were teaching IELTS preparation classes yet did not typically just adopt IELTS scoring without some modifications of their own. In doing this, however, we do not believe they should be regarded as in some way adopting a lazy, unmotivated or generally inferior approach to assessing the compositions, compared with that adopted in writing testing circles. Rather, as we will recapitulate below, there were ample signs of their adopting a thoughtful and strategic approach (cf. Chen and Bonner, 2017). Put another way, they evidenced perfectly reasonable beliefs underlying their practices.

Not only did teachers not adopt a standard scoring scale but they did not all feel the need to arrive at a final quantitative grade or score at all, even for their own purposes, let alone to be communicated to the writer. A qualitative rating in words rather than

numbers was often felt to be sufficient. Second, even if they did arrive at a score, although it shared with much professional exam rating of writing the characteristic of being analytic rather than holistic, it was arrived at much less explicitly (e.g. without any numerical calculation based on scores for more specific scales) and not even necessarily based on the same more specific criteria weighted in the same way for each essay. In performing this step, teachers viewed it as relevant to take into account individual writer factors, such as their past performance, and hence what aspects of their writing needed to be evaluated most. Such criteria would be deemed irrelevant, and indeed a source of unreliability or invalidity, in a professional testing environment.

Indeed, in the kind of assessment that I studied, a major influence at all points for most of the teachers was its locatedness in a specific context of individual students and their teaching and learning. There were many references, even before the composition was read, to what the student was learning at the time, what individual students or their peers had written in the past, and what it was important for the student to focus on for future success. These were clearly used to assist and inform the rating process. In a writing exam setting, however, scripts to be scored are usually anonymised, so examiners cannot use any knowledge they have of them individually even if the possess it. Furthermore, such examiners may not even know the general context of learning of the writers. Hence such knowledge could not come into play. Indeed, it is the precise intention of the precautions taken that they do not come into play as they are deemed undesirable as not consistent with objectivity. Yet where compositions are written and assessed for practice, with formative role, as part of the teaching, more than testing, process, such considerations are entirely appropriate, as also indicated by Lee (2011).

Finally, my findings showed that teachers assessing classroom practice writing took a further step. Beyond using contextual information to inform assessment, is to use the

information from assessment to inform the context, and there were indications of this also occurring. In other words, the teachers were using their assessment to provide themselves with pedagogical information, not just a measure of the writing ability of the student. This included updating their own knowledge about the writer, looking for reasons for what was written, maybe to suggest future topics for instruction, and of course offering feedback to the students, which in some cases seemed to be separated from the basic rating in the mind of the teacher, at other times not. None of these functions can be very readily performed by exam-like summative assessment.

5.3.2 Significance for teacher feedback research

My study also demonstrated the related yet distinct nature of research on teacher beliefs and practices concerning classroom writing assessment compared with those concerning teacher writing feedback.

Certainly, in our data there was a good deal of activity that could be considered as teacher feedback, both in terms of what the teachers wrote on the scripts and what they said that they would tell the writer (unprompted by the researcher, in the TA). However, what was shown was that this feedback forms part of a wider umbrella activity of teacher classroom writing assessment/rating whose other aspects are not usually elicited or reported in feedback studies, as when Bob says, for example "...*"It is agreed," – comma, common mistake. Must try and teach the class that one. "It is agreed that...."* ". Yet these other aspects form an important framework within which teacher feedback must be understood.

If for example a teacher in their feedback does not appear to be paying much attention to grammar errors, we need to understand this not only from the learners' point of view

(as feedback studies often explore) but also the teacher's. The teacher may be intending to mark all grammar errors but does not himself spot them all, or the teacher may be spotting them but deliberately not marking them for some reason. Only the wider perspective obtained with data on the assessment process such as we gathered can illuminate that, and in the latter case, the reasons the teacher had for their policy, such as an effect they wished to produce on the writer. Put another way, the wider perspective allows us to see where teachers are or are not making a distinction between their own assessment of a composition, and the one that they communicate to the writer, whether as a score or comments made on content or errors etc.

Furthermore, and connected with the preceding, only our wider perspective allows feedback to be seen within its full wider context of the teacher's beliefs and practices related to the assessment of classroom writing. This finally allows feedback to be seen in its proper perspective as just one of a range of uses that a teacher typically makes of information in student practice compositions. This we found to include maintaining an up to date perception of individual students' abilities and needs, and the gathering of information throwing light on past and perhaps future teaching. Without this perspective, it would be easy to get the idea from the writing literature that when a teacher reads a composition, his/her only purpose is or should be to provide feedback to the writer.

In this respect, I believe my study has added an important pedagogically-related dimension not fully recognised by the pioneering studies which inspired me such as Vaughan (1991), Lumley (2000) and Cumming et al. (2002). These kinds of findings, where connections are made between assessment and individual student backgrounds and pedagogy, are not highlighted in their models (see again Figure 2-4 and Tables 2-1 and 2-4). It is possible that these researchers simply overlooked some things that their

raters said that did not fit their prime focus on the rating/assessment process itself. Alternatively, perhaps it is the almost unparalleled authentic contextualisation achieved in the present study that allowed the appearance of such data from our teachers. Only when teacher assessment is elicited in relation to practice essays written by their own students in the context of a real course they are taking together can we hope to escape the limitations of the data obtained due to the unfamiliarity of teachers with teaching context and participants and gain access to that wider range of teacher cognitions which goes beyond the core of the rating itself. In sum, our study supports the view of writing researchers such as Connor-Linton (1995) concerning the value of research on the rating process. However, it does so not by showing how raters may affect test/examination rating and its validity and reliability, for instance by how they interpret the scoring rubric. Rather it illuminates the widely occurring kind of evaluation which teachers engage in which is not primarily exam related, but has teaching rather than testing purposes. It thus adds a strand to the broader study of teacher practices in the classroom, and teacher beliefs about these. We feel that this type of assessment needs to be better recognised and conceptualized as a distinct research area from testing, of which at present it is largely seen as an adjunct, and from feedback. While teacher classroom assessment of students remains perceived as a form of testing, it tends to be talked about in negative discourse, as if it is some deficient and biased form of true assessment (which is taken as meaning the assessment done by professional testers).

In this way, we support the still not fully implemented agenda laid out by McNamara (1997):

> "I am arguing that some of the most important research on language testing is not only technical; that is, research in language testing cannot consist only of a further burnishing of the already shiny chrome-plated quantitative armour of the

> language tester with his (too often his) sophisticated statistical tools and impressive n-size. Rather, I am arguing for the inclusion of another kind of research on language testing of a more fundamental kind, whose aim is to make us fully aware of the nature and significance of assessment as a social act." (p.460)

The difference is that we would prefer not to see this research as subsumed in the field of language testing at all but rather into that of language teaching, where features that from the testing point of view may be seen as deficiencies leading to bias (such as appealing to prior knowledge of the writer being rated) emerge as strengths, and allow us to illuminate not only the feedback which the teacher may then provide, but also a variety of other teacher classroom processes such as building perceptions of class members' abilities, and deciding on the focus of future teaching.

5.3.3 Implications

Classroom writing practice typically involves, or is of importance to, a number of different kinds of people. As suggested in chapter 1, our work may have some implications for all of them. We could sum up what we propose here as that all stakeholders in the language learning / teaching context need to be more aware of the different roles of teacher evaluation, their differing benefits, and how they can be achieved.

First, of course, the teachers who do the rating are central to the activity of classroom evaluation. For them the message of our work is similar to that for researchers, in that we would encourage them to become more aware of the distinction between evaluation which is for test/exam purposes, typically summative, everyday classroom evaluation, typically formative for the teacher, and the evaluation that is communicated to the

writers, typically formative for the students. It was apparent that such distinctions were not always clear in the minds of teachers, who in the TA and retrospective interview data sometimes showed, for example, a lack of clear distinction between what their personal 'true' evaluation was, and what they would communicate to the writer as feedback.

Teachers differed considerably in many areas which we investigated, such as: how far they invoked criteria or scoring based on professional tests; the kinds of data sources they drew on other than the scripts of the writers; and non-evaluative comments which they made. One message we have for them is that variation in such matters is acceptable and indeed desirable in classroom assessment. Along with this, however, comes the need to be more aware of what kinds of knowledge sources they are appealing to, and why, since we saw that a hallmark of our kind of rating was the use of multiple sources of information not usually available to testers in exam conditions, or, if available, prohibited from use by the rules of exam testing. Overall, there seems to be considerable scope for teachers to develop in their understanding of their own and each others' beliefs and practices. Since our study was descriptive and exploratory rather than evaluative, however, it must be made clear that we are not in a position to inform teachers definitively that this or that rating practice is good and another one bad.

Trainers, whether INSET or PRESET, may also benefit from our findings. They too need greater awareness of different types and roles of evaluation, the knowledge sources teachers draw upon, and so forth. In this way teacher professional development can be enhanced and, for example, not make the mistake of conflating teacher evaluation of students with testing, with the likely negative view taken of the former in that context. Training rather needs to enhance teacher understanding of the manifold

factors at work in their classroom evaluation, including its multiple purposes, information sources, types of comment that can be made, their possible functions, and so forth.

An issue that arises is how best to deliver training to teachers in such an area. We believe that a useful tool could be to employ as a training technique something like what we used as a research technique. It is possible that the very act of engaging in our research engendered some reflection by our participants on the matters we were focusing on, which would not have otherwise occurred. Although it was not part of the aim of our study to train the participants, possibly we not only obtained information from participants about their rating practices but also, incidentally, in the process made them more aware of what they were doing, so in effect served a pedagogical or training/teacher development purpose.

An additional approach could be that of encouraging peer discussion of rating beliefs and practices. As we noted earlier, there was considerable variation in what the participants in the present study actually did. Hence there is considerable scope for sharing of ideas and practices, treating teachers in a given context as a community of practice which benefits from peer communication and discussion to drive mutual development.

Arguably the area of our concern is one where training best takes the form of prompting reflection and discovery by teachers themselves, rather than traditional workshop presentations by a trainer which tend to promote one correct way of doing things. We have to admit, however, that training is increasingly associated with trying to get teachers to behave uniformly. In today's globalized world, where governments and educational institutions around the world are under great pressure to deliver a workforce

with adequate English to support a competitive economic position for a country in the international community, EFL teacher training has tended to become more top down and controlled. Countries decide on what they think is the best English policy to achieve their aims, including often starting to teach English at an earlier age, carefully selecting required textbooks etc. In this scenario teacher, professional development tends to take the form of teachers being trained to deliver a pre-decided model of TEFL education, rather than to self-develop. While this sort of model might suit training formal testers and examiners, where there exists an agreed body of wisdom on how to do things to achieve reliable and valid measures of proficiency, we do not feel that a strict form of practice can usefully be imposed on the sort of assessment which we are concerned with. It is likely to be highly context dependent. Nevertheless, there may be some place for common patterns that may facilitate consistency to be encouraged while the rationale for idiosyncratic behaviour can be closely investigated and appraised for its impact.

Course directors other than the teacher teaching a course, together with other relevant administrators, also need to become aware of what we described above. This includes managers of units such as the IA or CESC in our research context, or, in Saudi Arabia, the relevant deanship and departmental managers in the university hierarchy. There needs to be a clear distinction in everybody's mind between different types of evaluation. For instance, in the Saudi context there is an established tradition in schools of leaving most end of year language exam assessment to class teachers to devise and administer to their own classes. This would seem likely to result in poor rating for the exam purposes required, since exams require professionally made tests set and scored by trained testers. In the university context which we had in mind at the start of our work, there are two systems regarding the exams and assessment. The first system is

designed for the Preparatory year (between leaving school and embarking on majors) which is taught by competent English language teachers, and is assessed using a fixed rating scale. The second system is for English majors at the departmental level where, in the researcher's context, from year two, when students begin to be set genuine essay assignments, there is only one writing teacher, and she designs her own scale and criteria for both class practice writing and exam writing.

Course managers need not only to think carefully about what form of rating to use for exam/test purposes, but also be concerned with the sort of evaluation which we focused on, and with the evaluation which teachers relay to the students. These impact on day to day learning as much if not more than the former (which does so through backwash), and their effects on learners need to be understood. There needs to be movement towards forms of evaluation that are consistent with teachers' and managers' instructional goals and beliefs about the development of writing skills. It is thus expected that our findings provide evidence based insights on informal writing assessment by teachers, which can inform those who run the writing programmes in the contexts we researched, and may be informative beyond those contexts.

5.4 Limitations

A number of the limitations of the study, as we perceive them, relate to circumstances beyond our control.

First, of course, we were not able to investigate the rating of writing in the context which was of most interest to us from a practical pedagogical point of view. That is the context in which the researcher teaches, and which we would have liked our findings to be able to contribute directly to, i.e. the teaching of writing to English majors at a university in Saudi Arabia. As described in chapter 3, however, our pilot showed that

the teaching of writing at the level where multiple teachers were involved, and could be studied, was at such an elementary level that connected essays simply were not being written and assessed. In other words, true writing in the sense which we were concerned with was not occurring. At higher years of study where such writing was available, and being assessed, however, too few teachers were involved to sustain our research.

For that reason, we diverted to an available context in the UK. Although it was in some sense parallel, in being concerned with international students preparing to study at university through the medium of English, it differed in many ways from our originally chosen context, and hence its results cannot readily be generalised to that context. For one thing, the students were at a higher level of proficiency (typically IELTS 5) than those in KSA, and a mixture from a variety of countries, cultures and L1s, not classes with one L1, one culture and for the most part from one country, as found in KSA universities. Furthermore, they were studying in the country of the L2, so in effect a second language context rather than the foreign language context of KSA. Again, the teachers were different, in being mostly UK native speakers of English rather than the mixture found in KSA universities, which typically includes Arabic speaking teachers from a variety of countries and non-Arabic speakers from countries such as Pakistan or even Canada. Not only would the training backgrounds of such teachers probably be different from those we accessed, but also their general prior experience of teaching writing.

Secondly, even within our second context, and accessing two locations (the International Academy at the University of Essex, and the Colchester English Study Centre) we struggled to find enough participants willing to participate and, if initially willing, to then contribute as fully as we hoped. While we had intended to involve more

teachers, in the end we only obtained six, and even then, due to their pressure of work, it proved difficult to arrange enough sessions to cover all the data gathering we had hoped for. For instance, there was some unwillingness to perform TA while rating the number of scripts we would like. Furthermore, they did not always follow our instructions. For instance, one teacher read all the scripts before coming to the TA session, which was not what we instructed or wanted.

One consequence of the teacher limited willingness or availability was that there was some data which would have been interesting which we could not gather. For instance, following up on some of the literature reviewed in chapter 2, we had hoped to state and answer a research question concerning how teacher rating criteria and general process of handling rating did or did not change over time as they rated a succession of scripts. In practice, however, teachers were not able/willing to rate (with TA and follow up interview) even six scripts in one session. Hence, we had to abandon that aim, since changes of teacher rating behaviour over time as they rate multiple scripts is mainly of interest where they rate a whole set of scripts on one occasion.

Nevertheless, there were also some features of the study which were within our control and which, with hindsight, we would conduct differently if we were doing the study again. One is that we would attempt to make a clearer distinction between our efforts to access, on the one hand, teacher general retrospective thoughts about how they rate writing, using what scale and criteria and procedure etc., (in the general interviews), and, on the other, their reports of what they actually just did when rating a particular essay (in the immediate retrospective interviews). A better separation here would have allowed us to consider more clearly how far their beliefs, as stated in the former, did or did not chime with their practices, as seen in the latter. In fact, however, looking at our

interview transcripts, we can see that there were times when we allowed the scope to wander so that it is unclear whether what we were asking, or what was being given as an answer, really related to the teacher's general belief or to what they thought they just did in a specific instance when rating a particular script (i.e. their practice).

A final point is that, in retrospect, we feel we spent too much time bogged down in the procedure of attempting to code the TA data. This had the effect that we possibly paid less attention than should have been to the other qualitative data, which, although in some ways secondary in importance, was not as thoroughly coded and analysed. Furthermore, due to the long and effortful coding process, our impetus pushing the research forward weakened somewhat at the time when it was most needed, i.e. interpreting and writing up the results after the data analysis. Looking back, we should have exercised more rigorous control over the whole research process, e.g. by establishing from the start of the research a Gantt chart of realistic targets for certain dates, and adhering to it. It must be said however that changes of supervisor, together with the change of target location consequent upon the pilot results, described above, all contributed to making it difficult to keep to any ideal schedule.

5.5 Future research

The most obvious study still to be done is the one which we originally intended to do, within our home context. The problems which we encountered could perhaps be overcome if a researcher had access to more than one university. Then she would be able to access sufficient teachers teaching writing at a level where genuine essay texts were being produced by learners. In the same vein, many contexts around the world remain unstudied with respect to this important kind of writing assessment.

More widely, however, our study, being essentially exploratory, and undertaken in an under-researched area, leads the way towards many kinds of more focused studies of informal teacher rating of writing. For instance, it invites comparative studies where the same teachers are studied both when arriving at informal ratings of the type which we studied, and when scoring writing tests, done in exam conditions and with assigned standard marking schemes. This would allow a sharper analysis than we were able provide of just how much teachers change their rating habits in these two quite distinct situations, and how far the practices of one influence those of the other. In our study one teacher who had been trained as an IELTS writing examiner did evidence this in his informal rating. Still, a fuller study of teachers rating writing in different conditions, such as the testing versus the class practice condition, could illuminate much better how far such conditions are recognised as different by teachers, and how far transfer occurs between the two. In effect, such studies could be seen as rating strategy transfer studies in the same tradition that, for example, transfer of writing strategies between L1 and L2 writing by learners is studied.

Another more focused study would target further not the 'classroom rating - exam rating' interface, but the 'classroom rating - classroom feedback' interface, as described in chapter 1. Our study provided ample examples of teachers making a distinction between the rating they gave to a piece of writing, and voiced to the researcher, and the rating that they would communicate to the writer as part of feedback on the writing, although this was not always the case. The possible gap between the errors that the teacher notices and those that he/she chooses to underline or correct has widely been noted. The difference between the teacher's personally 'true' rating of writing and rating that is communicated to the learner, whether as a score or an evaluative word such as 'excellent', remains less researched. Yet this is a fascinating area, reliant as it is on a

wide range of beliefs that the teacher has about the learner, and strategies he/she may operate in order to achieve what he sees to be the optimum effect in terms of learner learning. Again, studies targeting just this aspect of rating and feedback are needed, examining closely what the teacher says and writes for learners in contrast with their personal internal ratings.

Finally, perhaps, a next step in research is to attempt to answer the key question in almost all research into teaching, which concerns effectiveness. While it is important to understand, what teachers do, and what they believe, as we have attempted to illuminate, in the end the purpose of an activity such as classroom writing practice is for students to improve. Thus, it remains the difficult but essential next step to try to evaluate what kinds of teacher informal rating practices, i.e. criteria, scales, information sources drawn upon, rating processes etc., are actually the most useful, and assist the teaching-learning enterprise best. This would connect with the feedback literature, which has to a considerable extent focused on this issue (2.2.3). As we have indicated above, we feel there can be no simple answer to this, since the answer will be context dependent at the narrow level of individual teacher and class even, in a particular institution and country. Hence answering such a question probably lends itself to highly localised research performed by practitioners themselves, with their classes, perhaps along the lines of the action research conducted by McPherron (2005) or Lee (2011).

REFERENCES

Afflerbach, P. P. (1990a). The influence of prior knowledge and text genre on readers' prediction strategies. *Journal of Reading Behavior*, *22*, 131-148.

Afflerbach, P. P. (1990b). The influence of prior knowledge on expert readers' main Idea construction strategies. *Reading Research Quarterly*, *25*(1), 31-46.

Afflerbach, P. P., & Johnston, P. H. (1984). Research methodology on the use of verbal reports in reading research. Journal of Reading Behavior, *16*(4), 306- 322.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. New York: Cambridge University Press.

Al-Lamki, N. (2009). *The beliefs and practices related to continuous professional development of teachers of English in Oman*. PhD Thesis, University of Leeds, UK.

Bachman, L. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

Bachman, L., Davidson, F., Ryan, K. & Choi, I. C. (1995). An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge–TOEFL Comparability Study. *Studies in Language Testing 1*. Cambridge: Cambridge University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-57.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Barkaoui, K. (2007a). Participants, texts, and processes in second language writing assessment: A narrative review of the literature. *The Canadian Modern Language Review*, *64*, 97-132.

Barkaoui, K. (2007b). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*, 86-107.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes*. Unpublished Ph.D. dissertation, University of Toronto, Toronto, Canada.

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, *44*(1), 31-57.

Barkaoui K. (2010b). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, *27*(4), 515-535.

Barkaoui, K. (2010c). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54-74.

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

Baumann, J. F., Jones, L. A., & Seifert-Kessell, N. (1993). Using think-alouds to Enhance children's comprehension monitoring abilities. *The Reading Teacher*, *47*(3), 184-193.

Bonner, S., Torres Rivera, C., & Chen, P. (2018). Standards and assessment: Coherence from the teacher's perspective. *Educational Assessment, Evaluation and Accountability*, 30(1), 71-92.

Borg, M. (2001). Key concepts in ELT. Teachers' beliefs. *ELT Journal*, 55(2), 186-188

Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching*, 36(2), 81-109.

Borg, S. (2006). *Teacher cognition and language education: Research and practice*. London: Continuum.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. London and New York: Routledge.

Brindley, G. (1998c). Describing language development? Rating scales and second language acquisition. In L. Bachman, & A. D. Cohen, (Eds.), *Interfaces*

between second language acquisition and language testing research, 112-140. Cambridge: Cambridge University Press.

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

Brown, J. D., & Rodgers, T. S. (2002). *Doing second language research: An introduction to the theory and practice of second language research for graduate/Master's students in TESOL and Applied Linguistics, and others*. Oxford: Oxford University Press.

Bukta, K. (2007). *Processes and outcomes in L2 English written performance assessment: Raters' decision-making processes and awarded scores in rating Hungarian EFL learners' compositions*. Unpublished Ph.D. dissertation, University of Pécs, Hungary.

Calderhead, J. (1996). Teachers: Beliefs and knowledge. In D. Berliner & R. Calfee (Eds). *Handbook of Educational Psychology* (pp.709-725). New York: Macmillan.

Callear, D., Jerrams-Smith, J., Victor, S. (2001). Bridging gaps in computerised assessment of texts. *In Proceedings of the IEEE International Conference on Advanced Learning Techniques (ICALT)*, 139-140. Madison, Wisconsin: IEEE.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, *18*, 65-81.

Chen, P. & Bonner, S. (2017). Teachers' beliefs about grading practices and a constructivist approach to teaching. *Educational Assessment*, 22(1), p18-34.

Chinda, B. (2009). *Professional development in language testing and assessment: A case study of supporting change in assessment practice in in-service EFL teachers in Thailand*. Ph.D. dissertation, University of Nottingham, Nottingham, UK.

Cohen, A. D. (1987). Using verbal reports in research on language learning. In C. Faerch, & G. Kasper, (Eds.), *Introspection in second language research*, *30*, 82-95.

Clevedon, England: Multilingual Matters.

Cohen, A.D., & Cavalcanti, M.C. (1990). Feedback on compositions: Teacher and student verbal reports. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom*, 155-177. Cambridge: Cambridge University Press.

Cohen, A. D. (1994). *Assessing language ability in the classroom* (2nd Ed.). Boston: Heinle & Heinle.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*(2), 163-178.

Connor-Linton, J. (1995): looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly*, *29*(4), 762-765.

Connor, U., & Carrell, P. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspective*, 141-160. Boston: Heinle and Heinle.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning*, *39*, 81-141.

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

Cumming, A. (1997). The testing of writing in a second language. In C. Clapham & D. Corson, (Eds.), *Encyclopedia of language and education: Language testing and assessment*, *7*, 51-65. Dordrecht, Netherlands: Kluwer.

Cumming, A., & Riazi, A. (2000). Building models of adult second language writing instruction. *Learning and Instruction, 10*(1), 55-71.

Cumming, A. H., Kantor, R., & Powers, D. E. (2001). Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework. *TOFEL Monograph Series, MS-22*. Princeton, NJ: Educational Testing Service.

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

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: UCLES, Cambridge University Press.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, *5*, 7-29.

Douglas, D., & Myers R. (2000). Assessing the communication skills of veterinary students: whose criteria?. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquim, Orlando, Florida* 60-79. Cambridge: Cambridge University Press

Douglas, D. (2010). *Understanding language testing*. London, UK: Hodder Education.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford: Oxford University Press.

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Daughty & M. H. Long (Eds.), *The handbook of second language acquisition*, 589- 630. Malden, MA: Blackwell.

Dörnyei, Z., & Csizer, K. (2012). How to design and analyze surveys in second language acquisition research. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition: A practical guide,* 74-94. Malden, MA: Wiley-Blackwell.

Eckes, T. (2005b). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221.

Eckes, T. (2008): Raters types in writing performance assessment: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition

with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112.

Engelhard, G., Jr., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model. *College Board Research Report (No. 2003-1)*. New York: College Entrance Examination Board.

Erdosy, M. U. (2000). *Exploring the establishment of scoring criteria for writing ability in a second language: The influence of background factors on variability in the decision- making process of four experienced raters of ESL compositions*. Unpublished MA Thesis, OISE, University of Toronto, Toronto, Canada.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions (ETS-TOEFL Research Report No. 70, RR 03-13)*. Princeton, NJ: Educational Testing Service.

Ericsson, K. A. (2002). Towards a procedure for eliciting verbal expression of non-verbal experience without reactivity: Interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology*, *16*(8), 981- 987.

Ericsson, K., & Simon, H. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Ericsson, K. & Simon, H. (1987). Verbal reports on thinking. In C. Færch and G. Kasper (Eds.), *Introspection in Second Language Research*, 24-53. Clevedon, England: Multilingual Matters.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data* (Revised Ed.). Cambridge, Mass: MIT Press.

Fang, Z. (1996). A review of research on teacher beliefs and practices. *Educational Research*, 38(1), 47-66.

Fathman, A. K., & Whalley, E. (1990). Teacher response to student writing: focus on form versus content. In B. Kroll (Ed.), *Second Language Writing: Research*

Insights for the Classroom, 178- 190. Cambridge: Cambridge University Press.

Ferris, D. R. (1995). Student reactions to teacher response in multiple-draft composition classrooms. *TESOL Quarterly*, *29*, 33-53.

Ferris, D. (2006). *Responding to student writing: Approach, response, follow-up, & evaluation*. 3rd Annual Symposium on Multilingual Student Writers: UC Berkeley College Writing Programs.

Ferris, D.R. & Hedgcock, J.S. (2005). *Teaching ESL composition: Purpose, process and practice* (2nd edition). Mahwah, NJ: Lawrence Erlbaum.

Flower, L., & Hayes, J. R. (1981). A Cognitive process theory of writing. *College composition and communication*, *32*(4), 365-387.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods*, 75–98. New York: Longman.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Gibson, W., & Brown, A. (2009). *Working with qualitative data*. Los Angeles: Sage Publications.

Ghanbari, B., Barati, H., & Moinzadeh, A. (2012). Rating scales revisited: EFL writing assessment context of Iran under scrutiny. *Language Testing in Asia*, *2*(1), 83-100.

Green, A. J. K. (1997). *Verbal protocol analysis in language teaching research*. Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.

Hamp-Lyons, L. (1987). *Testing second language writing in academic settings*. Unpublished Ph.D dissertation. University of Edinburgh, Scotland.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In B. Kroll (Ed.) *Second language writing: Research insights for the classroom*, 69-87. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, 241-276. Norwood, NJ: Ablex.

Hayes, J. R., Flower, L. (1981). *Uncovering cognitive processes in writing: An Introduction to Protocol Analysis*. ERIC Clearinghouse.

Henderson, D., Rupley, W., Nichols, J., Nichols, W. & Rasinski, T. (2018). Triangulating teacher perception, classroom observations, and student work to evaluate secondary writing programs. *Reading & Writing Quarterly*, 34(1), 63-78.

Hesse-Biber, S., & Leavy, P. (2006 2010): *The Practice of Qualitative Research*. SAGE.

Hilden, K., & Pressley, M. (2004). Verbal protocols of reading. In N. K. Duke & M. H. Mallette (Eds.), *Literacy Research Methodologies*, 308-321. New York and London: Guilford Press.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, *18*(1), 87-107.

Huang, J. (2010). Factors affecting the assessment of ESL students' writing. International *Journal of Applied Educational Studies*, *5*(1), 1-17.

Huot, B. A. (1988). *The validity of holistic scoring: A comparison of talk-aloud protocols of expert and novice holistic raters*. Unpublished Ph.D. dissertation, Indiana University of Pennsylvania, US.

Huot, B. A. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201-213.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*, 206-236. Cresskill, NJ: Hampton Press.

Hyland, F. (2003). Focusing on form: Student engagement with teacher feedback. *System*, *31*(2), 217-230.

Hyland K. (2003). Genre-based pedagogies: A social response to process. *Journal of Second Language Writing*, *12*(1), 17-29.

Hyland, K. (2003). *Second language writing. New York*, Cambridge University Press.

Hyland, K. & Hyland, F. (2006). State of the Art article: Feedback on second language students' writing. *Language Teaching*, 39(02), 83-101.

James, C. (1998). *Errors in Language Learning and Use*. New York: Routledge.

Janopoulos, M. (1992). University faculty tolerance of NS and NNS writing errors: A comparison. *Journal of Second Language Writing*, *1*(2), 109-121.

Janopoulos, M. (1993). Comprehension, communicative competence, and construct validity: holistic scoring from an ESL perspective. In M. A. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment*, 303-325. Cresskill: Hampton Press.

Janopoulos, M. (1995). Writing across the curriculum, writing proficiency exams, and the NNS college student. *Journal of Second Language Writing*, *4*, 43-50.

Kagan, D. M. (1990). Ways of evaluating teacher cognition: Inferences concerning the Goldilocks Principle. *Review of Educational Research Association*, 60(3), 419-469.

Khongput, S. (2010). EFL Writing Assessment Practices: Teachers' Perspectives. *In The 36th International Association for Educational Assessment (IAEA) Annual Conference*, Bangkok Thailand.

Khongput, S. (2014): *English language writing assessment: Teacher practices in Thai universities.* Unpublished PhD dissertation, University of New South Wales, Australia.

Kinay, I. (2018). Investigation of prospective teachers' beliefs towards authentic assessment. *World Journal of Education*, 8(1), 75-85.

Kinay, I. & Ardiç, T. (2017). Investigating teacher candidates' beliefs about

standardized testing. *Universal Journal of Educational Research*, 5(12), 2286-2293.

Klingner, J. K. (2004). Assessing reading comprehension. *Assessment of Effective Intervention, 29*(4), 59-70.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing, 28*(2), 179–200.

Koda, K. (2005). Insights into second language reading. Cambridge: Cambridge University Press.

Kobayashi, T. (1992). Native and non-native reactions to ESL compositions. *TESOL Quarterly, 26*(1), 81-112.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219-240.

Lee, N. S. (2009). Written peer feedback by EFL students: Praise, criticism and suggestion. Komaba. *Journal of English Education, 1*, 129-139.

Lee, I. (2009). A new look at an old problem: How teachers can liberate themselves from the drudgery of marking student writing. *Prospect: An Australian Journal of Teaching/Teachers of English to Speakers of Other Languages (TESOL), 24*(2), 34-41.

Lee, I. (2011). Formative assessment in EFL writing: An exploratory case study. *Changing English: Studies in Culture and Education*, 18(1), 99-111.

Leki, I. (1991). The preferences of ESL students for error correction in college-level writing classes. *Foreign Language Annals, 24*, 203-218.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: the issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition, 26*(1), 35-57.

Lindsey, P., & Crusan, D. (2011). How faculty attitudes and expectations toward student nationality affect writing assessment. *Across the Disciplines, 8*(4). Retrieved September 17, 2014.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), Evaluating writing: *Describing, measuring, judging*, 33-66. Urbana, IL: National Council of Teachers of English.

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

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

Lumley, T. (2005): *Assessing second language writing: The rater's perspective*. Frankfurt Germany: Peter Lang.

Lumley, T., & Brown, A. (2005). Research methods in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*, 833-855. Mahwah, NJ: Lawrence Erlbaum.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design.* Mahwah, NJ: Lawrence Erlbaum.

Matsumoto, K. (1994). Introspection, verbal reports and second language learning strategy research. *The Canadian Modern Language Review, 50* (2), 363-385.

Mcgillicuddy-De Lisi, A.V., & De Lisi, R. (2001). *Biology, society, and behaviour: The development of sex differences in cognition*. Greenwich, CT: Ablex.

McNamara, T. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466.

McNamara, T. F. (1996). *Measuring second language performance*. London: Addison Wesley Longman.

McNiff , J. (1993). *Action research: Principles and practice*. NY: Routledge.

McPherron, P. (2005). Assumptions in assessment: the role of the teacher in evaluating ESL students. *The CATESOL Journal, 17*(1), 38-54.

Milanovic M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composition marker. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language*

Testing Research Colloquim, Cambridge and Arnhem, 3, 92-114. Cambridge: Cambridge University Press and University of Cambridge local Examinations Syndicate.

Mullis, I. V. (1984). Scoring direct writing assessments: what are the alternatives? *Educational Measurement: Issues Practice, 3*(1), 16-18.

Nakamura, Y., (2002). A comparison of holistic and analytic scoring methods in assessment of writing. *Paper presented at the Interface between Interlanguage, Pragmatics and Assessment: Proceedings of the 3rd annual JALTA Pan-SIG Conference*. Tokyo, Japan: Tokyo Keizai University.

Nunan, D. (2012). *Research methods in language learning*. Cambridge: Cambridge University Press.

O'Loughlin, K. J. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics, 17*(1), 23-44.

Olshavsky, J. E. (1976-1977). Reading as problem-solving: An investigation of strategies. *Reading Research Quarterly, 12* (4), 654-674.

Olson, G. M., Duffy, S. A., & Mack, R. L. (1984). Think-out-loud as a method for studying real-time comprehension processes. In D. E. Kieras & M. A. Just. (Eds.), *New Methods in Reading Comprehension Research*, 253-286. Hillsdale: Lawrence Erlbaum Associates.

Paltridge, B. (1994). Genre analysis and the identification of textual boundaries**.** *Applied Linguistics, 15*(3), 288-299.

Perkins, K. (1983). On the use of composition rating techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17*(4), 651-671.

Pit Corder, S. (1975). *Introducing Applied Linguistics*. UK: Penguin Books Ltd.

Pressley, M., & Afflerbach, P. (1995). *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pritchard, R. (1990a). The effects of cultural schemata on reading and processing strategies. *Reading Research Quarterly, 25* (4), 273-295.

Pritchard, R. (1990b). The evolution of introspective methodology and its implications for studying the reading process. *Reading Psychology: An International Quarterly, 11*, 1-13.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M.M. Williamson, & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*, 237-265. Cresskill, NJ: Hampton Press.

Purves, A. C. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English*, 108-122.

Rafoth, B. A., & Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication, 1*(4), 446-458.

Raimes, A. (1990). The TOEFL test of written English: causes for concern. *TESOL Quarterly 24*(3), 427-442.

Richardson, V. (1996). The role of attitudes and beliefs in learning to teach. In J. Sikula (Ed.), *Handbook of research on teacher education* (2nd ed., pp. 102–119). New York, NY: Macmillan.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquim*, Orlando, Florida, 129-152. Cambridge: Cambridge University Press.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors.* Unpublished Ph.D. dissertation, University of Toronto, Toronto, Canada.

Santos, T. (1988). Professors' reactions to the writing of nonnative-speaking students. *TESOL Quarterly, 22*(1), 69-90.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*, 1-30.

Shaw, S. (2001). Issues in the assessment of second language writing. *Research Notes*, 6(1), 2-6.

Shaw, S. D., & Weir, C. J. (2007). Examining writing: Research and practice in assessing second language writing. *Studies in Language Testing, 26*. Cambridge, UK: UCLES/Cambridge University Press.

Shi, L. (2001). Native- and non-native-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

Shirazi, M. A. (2012). When raters talk, rubric fall silent. *Language Testing in Asia, 2*(4), 123-139.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics, 15*, 188-211.

Smagorinsky, P. (Ed.). (1994). Speaking about writing: Reflections on research methodology. Thousand Oaks, CA: Sage.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment, 1*, 159–189. Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Someren, M. V., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: a practical approach to modelling cognitive processes.* London: Academic Press.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*, 163-182.

Tedick, D. J., & Mathison, M. A. (1995). Holistic scoring in ESL writing assessment: What does an analysis of rhetorical features reveal. *Academic writing in a second language: Essays on research and pedagogy*, 205-230.

Torrance, H. (1998): Learning from research in assessment: A response to Writing assessment. Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 31-37.

Troia, G. A. & Graham, S. (2016). Common core writing and language standards and aligned state assessments: A national survey of teacher beliefs and attitudes. *Reading and Writing: An Interdisciplinary Journal*, 29(9), 1719-1743.

Urquhart, S. & Weir, C. J. (1998). *Reading in a Second Language: Process, Product and Practice*. London: Addison Wesley Longman Ltd.

Vann, R., Lorenz, F., & Meyer, D. (1991). Error gravity: Faculty response to errors in the written discourse of non-native speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, 181-195. Norwood, NJ: Ablex.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, 111-125. Norwood, NJ: Ablex.

Wade, S. E. (1990). Using think alouds to assess comprehension. *The Reading Teacher, 43*(7), 442-451.

Wade, S. E., Trathen, W. & Schraw, G. (1990). An analysis of spontaneous study strategies. *Reading Research Quarterly, 25*(2), 147-166.

Wallace, M.J. (1991). *Training foreign language teachers: A reflective approach.* Cambridge: Cambridge University Press.

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

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

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

White, C., Schramm, K., & Chamot, A. U. (2007). Research methods in strategy research: Re-examining the toolbox. In A. D. Cohen and E. Macaro (Eds.), *Language Learner Strategies: Thirty Years of Research and Practice, 30*, 93-116. Oxford: Oxford University Press.

William, D. (1994). Assessing authentic tasks: alternatives to mark-schemes. *Nordic Studies in Mathematics Education, 2*(1), 48-68.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4,* 83-106.

Wolfe, E. W., Kao, C. & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication, 15*, 465-492.

Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment, 2*, 37-56.

Wood, R. (1993). *Assessment and testing: a survey of research*. Cambridge: Cambridge University Press.

Wu, S. M. (2010). Investigating raters' use of analytic descriptors in assessing writing. *Reflections in English Language Teaching*, *9*(2), 69–104.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing*, *15*(3), 179-200.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly*, 19(1), 79–97.

Zhang, W. (1998). *The rhetorical patterns found in Chinese EFL student writers' examination essays in English and the influence of these patterns on rater*

response. Unpublished Ph.D. dissertation, Hong Kong Polytechnic University.

APPENDICES

Appendix A: Some empirical studies which elicited TAPs to investigate the essay rating process, in chronological order (adapted from Barkaoui, 2010, p. 75)

Study	Purpose	Raters	Essays	Rating scale	Justification for use of TAPs	Empirical evidence on TA effect	Cautions on TAP limitations	Discussion of rater variation in TAPs	Other TAP-related comments
Vaughan (1990)	Descriptive (rating process and criteria)	9 experienced raters	6 borderline essays by native and ESL students	Holistic with 6 levels	TA used in research on writing; Ericsson and Simon (1980) on non-reactivity of TA	NO	NO	NA	
Cumming (1990)	Comparative (experienced and novice raters; writers differing in proficiency and in writing experience)	6 experienced and 7 novice raters	12 ESL essays	Analytic with 3 rating dimensions and 4 levels	TAPs provide more directly valid information about rating than score analysis; Ericson & Simon (1984)	NO	Yes: both incompleteness and reactivity	NO	
Huot (1993)	Comparative (novice and experienced raters with and without holistic scoring)	4 English teachers with holistic rating experience and training and 4 without	42 essays by native students	Holistic and with no rating scale	Ericsson and Simon (1984); Previous studies using TAPs	NO	NO	NO	Used multiple methods to address TAP incompleteness
Connor and Carrell (1993)	Descriptive (how raters interpret essay tasks and use of rating scale)	5 experienced trained raters	5 ESL essays	Holistic with 6 levels	TA used in psychology; TAPs used in previous ESL writing and rating research	NO	NO	NA	
Weigle (1994)	Comparative (rating process before and after training)	4 experienced raters	4 ESL essays	Analytic with 10 levels and 3 rating dimensions	Reference to Ericsson and Simon (1993); Previous studies using TAPS	NO	Yes incompleteness; but not reactivity	Yes, amount of verbalization varies across raters	
Milanovic et al. (1997)	Comparative (4 rater groups with varying experiences)	16 raters in 4 groups in terms of experience	40 ESL essays	Holistic with 6 levels and 2 components	NO	NO	NO	NO	Used multiple methods to address TAP incompleteness
Delaruelle (1997)	Comparative (rating process across text types and rater experience)	3 experienced and 3 inexperience raters	36 ESL essay on interpersonal and persuasive tasks	Analytic with 5 levels and 4 rating dimensions	Ericsson and Simon (1984); Previous research used TAPs	NO	NO	NO	
Wolfe (1997)	Comparative (raters with different levels of scoring proficiency)	3 groups of 12 raters with different scoring proficiency	24 essays	Holistic with 6 levels	Ericsson and Simon (1993); Previous research using TAPs	NO	NO	NO	Argue TAPs can be used in rater selection, training and monitoring

Study	Purpose	Raters	Essays	Rating scale	Justification for use of TAPs	Empirical evidence on TA effect	Cautions on TAP limitations	Discussion of rater variation in TAPs	Other TAP-related comments
Wolfe et al. (1998)	Comparative (raters with different levels of scoring prof)	3 groups of 12 raters with different scoring proficiencies	24 essays	Not specified	Ericsson and Simon (1993); Previous research using TAPs	NO	NO	NO	Argue TAPs can be used in rater selection, training and monitoring.
DeRemer (1998)	Descriptive (rating process, how raters define rating task)	3 experienced raters	24 essays by 7 native students	Analytic with 4 levels and 5 rating dimensions	Ericsson and Simon (1993); Previous studies using TAPs	NO	Yes, incompleteness; but not reactivity	NA	
Weigle (1999)	Comparative (novice and experienced raters; 2 prompts; before and after training)	8 novices and 8 experienced raters	60 ESL essays on 2 prompts	Analytic with 10 levels and 3 rating dimension	Ericsson and Simon (1980,1984); Previous studies using TAPs.	NO	Yes, incompleteness; but not reactivity	NO	
Saki (2000)	Descriptive (to build a model of the essay rating process)	6 experienced ESL raters	12 essays by first- year university students	Holistic with 5 levels	TAPs provide rich information on rater behaviour; TAPs used in previous research successfully	NO	NO	NA	
Smith (2000)	Descriptive (raters' reading strategies and use of rating criteria)	6 experienced trained raters	3 ESL essays	Analytic with 6 rating dimensions	Ericsson and Simon (1980); Previous research using TAPs	NO	Yes, incompleteness but not reactivity	NA	
Cumming et al. (2002; 2001)	Descriptive and comparative (to develop model of rating process; to compare process across tasks and raters)	Various numbers of ESL and English composition teachers in different phases of study 4 raters form various backgrounds	Various numbers of ESL essays at different phases of study	Impressionistic with 6 levels and no descriptors	TAP used in previous studies	NO	Yes, both incompleteness and reactivity	NO	Multi-phase study, with different raters and tasks, to check and confirm findings from TAPs
Erdosy (2004)	Descriptive and comparative (effects of raters' background on rating process and criteria)	4 raters from various backgrounds	60 ESL essays	Impressionistic with 6 levels and no descriptors	Insights TAPs provide. Previous studies using TAPs	Yes: Quotes from one participant about TAP reactivity and incompleteness	Yes, both incompleteness and reactivity	Yes, TA completeness varies across individuals	Used multiple methods to triangulate data and address limits of TAPs

Study	Purpose	Raters	Essays	Rating scale	Justification for use of TAPs	Empirical evidence on TA effect	Cautions on TAP limitations	Discussion of rater variation in TAPs	Other TAP-related comments
Lumley (2002,2005)	Descriptive (to develop a model of the rating process and role of the rater and rating scale in the process)	4 motivated, articulated and highly trained and experienced ESL raters	24 ESL essays	Analytic	Extensive critical review of literature on TAPs	Yes, TAPs show veridicality and reactivity using FACETS score analysis and interviews with participants about TAPs	Yes, both incompleteness and reactivity	NO	Provides a critical review of research using TAPs.

Appendix B: The consent form

UNIVERSITY OF ESSEX

FORM OF CONSENT TO TAKE PART IN A RESEARCH PROJECT

CONFIDENTIAL

Title of the project

Teachers' ESL writing assessment practices

Name of principal investigator:

Manal Alghannam.

What is the project about?

You are invited to participate in a study of writing assessment practices of ESL writing teachers. The purpose of this is to explore teachers' actual practices of marking their students' writing compositions/essays. The study will focus on the marking process and teachers' decision making while marking the compositions/essays. Information you provide will be made anonymous and your participation is voluntary. You can decide to withdraw from taking part in this research at any time without giving your reasons for doing so.

Taking Part: please Tick YES/NO in the boxes:

I have read and understood the project information given above. { }

I have been given the opportunity to ask questions about the project. { }

Underline where appropriate

I agree/disagree to take part in the project. Taking part in the project will include interview, thinking aloud while evaluating your students' compositions as well as retrospective interview, with all of them being audio-recorded.

I understand that my taking part is voluntary; I can withdraw from the study at any time and I do not have to give any reasons for why I no longer want to take part.

Use of the information I provide for this project

I understand my personal details such as name, email address and phone number will not be revealed to people outside the project.

I understand that my words may be quoted in publications, reports, web pages, and other research outputs (although my identity would be disguised).

Use of the information I provide beyond this project

I agree/disagree for the data I provide to be archived at the UK Data Archive.

I understand that other genuine researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.

I understand that other genuine researchers may use my words in publications, reports, webpages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.

I agree/disagree with everything stated above.

_____	_____	_____
Name of participant [printed]	Signature	Date

_____	_____	_____
Researcher [printed]	Signature	Date

Appendix C: General interview questions

I. Background information

1. Can you tell me please first about yourself?
2. What is your nationality?
3. What is your first language? Do you speak another language?
4. What is your highest level of education? Can you please say where you completed your degree?
5. What is your first language?
6. Other languages you have a command of?

II. Teaching experience

1. How long have you been teaching English? (probe questions)
2. How long have you been teaching in the International Academy/CESC?
3. How long have you been teaching writing?
4. How long have you been teaching writing for EFL students in countries other than England?

III. Training experience in marking writing

1. What is your experience in assessing/evaluating/correcting English writing compositions/essays?
2. Or, Can you tell me please how do you describe yourself or your skill in assessing/evaluating/correcting English writing compositions/essays? I mean do you consider yourself expert, competent or novice?
3. Have you ever had any training regarding marking writing? ('Training' includes practice in marking writing and advice about how to mark writing)
4. How would you utilize your past experience as an expert/novice, if you consider yourself an expert/ a novice? I mean how does your experience impact on your assessment?

IV. Views on good writing

1. Can you tell me please what makes a good piece of writing from your point of view? Offered categories for them to rate on flash cards
2. What are the most important aspects of good writing?
3. How do you find other ways to assess your students' writing? I mean can you tell me from where you obtain your methods of assessment, and how do you use them?
4. What criteria do you use in marking your students' writing? I will show some cards and please show which one indicates you.
    - No criteria
    - University's criteria (the central criteria that your **university** gives to all university teachers to apply in **any** writing courses)
    - Department's criteria (the central criteria that your **department** gives to all teachers in the **department** to apply in **any** writing courses)
    - Course criteria (criteria that your department gives to the teachers teaching **a particular** writing course)
    - Your own criteria (the criteria that **you and/or you and your co-teachers have created** to use in your writing course)
5. What are the most important factors that you look at when assessing students' writing?
6. Or what are the qualities that you look at when assessing/evaluating students writing?

V. Classroom practices regarding writing assessment

1. How do you usually inform your students about your methods of assessment/ evaluation/correction or rating? I will show you some cards and I want you to show which card represents you:
   To what extent do you agree with Reem? Why?
   To what extent do you agree with Hala? Why?
   Reem: "I never say anything to my students about how I assess their writing. I don't think they need to know. And I don't think they would be interested."
   Hala: "I tell my students the criteria I use to assess their writing because I think it motivates them to try to make sure their writing is satisfactory according to my criteria."
2. Can you tell me how the students know what makes a good piece of writing?
3. Does the curriculum guide you to any particular framework or rating scale to follow?
4. If you cooperate with a co-teacher, how do you share your rating criteria? If not, can you give more details how you validate your scoring system to make sure that all students have equal evaluation?
5. How often do you assess your students? Can you tell what means you use? Like portfolio assessment. etc.

VI. General issues

1. How do you describe the time/process of rating/correcting students' compositions?
2. Can you please tell me what are the challenges that you experience in writing assessment/evaluating/rating?
3. How to overcome these challenges?
4. From your experience, what are the factors that affect your decisions?
5. What are the difficulties that you may find during rating?
6. How do you usually act when you encounter such difficulties in the rating process?

VII. Reactions to the task of producing TA protocols

1. Did you find it easy or difficult; natural or unnatural; time-consuming or not?
2. Do you feel it affected the way in which you rated? If so, how?
3. Do you have any comments at all to make about the compositions you rated today (whatever time)? Were they typical students' compositions?

VIII. Conclusion

1. Is there anything else at all you would like to discuss?
2. What about the test?
3. What about curriculum design?
4. Or about what I have asked you to do?

Appendix D: The verbal report instructions

Instructions:

In this research, I am interested in what you think about as you carry out the tasks I am going to give you. To do this, I am going to ask you to think aloud as you work on a task. By 'think aloud' I mean that I want you to say out loud everything that are you thinking from the time you start the task until you complete it. I would you like to talk constantly from the time you commence the task until you have completed it. It is important that you do not plan out or try to explain to me what you are thinking. It may help if you imagine that you are in a room by yourself.

It is very important that you keep talking. If you are silent for any period of time, I shall remind you to keep talking.

- Do you understand what I am asking you to do? Do you have any questions?
- We shall start with a few practice problems. This is to help you get used to thinking aloud and saying everything that is on your mind.
- When you have finished practicing and are used to thinking aloud we will then move on to the real task about assessing writing.

Appendix E: The TA training tasks

Instructions:

In this research, I am interested in what you think about as you carry out the tasks I am going to give you. To do this, I am going to ask you to think aloud as you work on a task. By 'think aloud' I mean that I want you to say out loud everything that are you thinking from the time you start the task until you complete it. I would you like to talk constantly from the time you commence the task until you have completed it. It is important that you do not plan out or try to explain to me what you are thinking. It may help if you imagine that you are in a room by yourself.

It is very important that you keep talking. If you are silent for any period of time, I shall remind you to keep talking.

- Do you understand what I am asking you to do? Do you have any questions?
- We shall start with a few practice problems. This is to help you get used to thinking aloud and saying everything that is on your mind.
- When you have finished practicing and are used to thinking aloud we will then move on to the real task about assessing writing.

I would like you to think aloud as you add up all the windows in your house.

- Now I would you to tell me what you can remember about what you were thinking from the time you read the practice question until you gave your answer. I am interested in what you can actually remember, not what you think you may or should have thought. If possible, it would be best if you can tell what you remember in the order in which the memories occurred as you worked through the question. If you are not sure about any of your memories, please say so. I do not want you to try to solve the problem again, I just want you to tell me what you can remember thinking. Now tell me what you can remember.

Appendix F: The first pilot study warm up task (the arithmetic task)

- I would like you to complete some more practice problems. I would like you to do the same thing for each of these problems. I want you think aloud as before as you work on the problems and once you have finished the problem, I want you to tell me all that you can remember about your thinking.

Do you have any questions?

First, I would like you to think aloud as you multiply two numbers in your head. The numbers are 25 and 15.

- Now I would you to tell me what you can remember about what you were thinking from the time you read the practice question until you gave your answer. I am interested in what you can actually remember, not what you think you may or should have thought. If possible, it would be best if you can tell what you remember in the order in which the memories occurred as you worked through the question. If you are not sure about any of your memories, please say so. I do not want you to try to solve the problem again, I just want you to tell me what you can remember thinking. Now tell me what you can remember.

Appendix G: The second pilot warm up task (the verbal task)

- I would like you to solve an anagram. I will show you a series of letters. Your task is to unscramble all the letters and to rearrange them to form a word in English. For instance, the letters ODOR can be rearrange to form the word DOOR.

  Please talk aloud as you work on the following anagrams.

  The anagram is: tsdsarie

  The anagram is: drenag

- Now tell me all that you can remember about your thinking.

- Now I would like you to imagine that you want to go to the supermarket or the shopping mall. I want you to think aloud while you are writing a list of words.

The task is to write your shopping list beginning from your most wanted items.

- Now tell me all that you can remember about your thinking.

Appendix H: The researcher observation sheet.

Researcher observation sheet for the think aloud reporting

Appendix I: The symbols used for the transcription of the think-aloud protocols and retrospective interviews

Symbol	Meaning
( )	for uncertain transcription
x	incomprehensible item, one word only
xx	incomprehensible item of phrase length
xxx	incomprehensible item beyond phrase length
…	three dots indicate a pause of five seconds or more
-	a hyphen indicates an incomplete word (e.g., wait plea-)
---	three hyphens indicate an interrupted or incomplete sentence.
[ ]	square brackets indicate a word or phrase in a language other than English, also naming the language (e.g., [Arabic])
“ ”	quotation marks indicate text read directly from the original composition
___	underline word or sentence if it is read from another source.

Adapted from Cumming et al. (2002)

Appendix J: An example of a complete transcription of a think-aloud protocol

Transcription template
Transcript's number[7]: 1/ LEVEL[8]: C1B
Teacher's ID: IA /Z[9]
Booking ID: 99729[10]
Event ID: 99729
Date: 20-02-2014
Day: Thursday (week21)
Room: 3.407

Recording time: **1:05:11 Task: 1 Start of the task: 00:16 End of the task: 18:58 Real time: 19 minutes**

	Time	Speaker	Content	Observation	Notes
The beginning of the task	0:16	Rater 1	What was the task, the task was, um "in many countries", that is the second task of ELTS, "children are engaged in some kind of paid work, some people regard this as completely wrong, while others consider it as work experience xx	He is reading the task out loud	He started without any special approach. He immediately took the first paper at the top without rearranging them.
	0:45	Rater 1	What I expect from the students is to write two paragraphs. One with and one against and one their		

[7] This refers to student's paper sheet.
[8] This refers to the level of the class.
[9] This refers to the rater, where IA is the institution the rater works for(International Academy). Z is the pseudonym of the rater.
[10] This session is done in quite room provided by the Lang & Ling department after it has been arranged and booked. This number refer to the booking.

			opinion and conclusion and introduction. So probably five paragraphs.		
	0:53	Rater 1	One, two, three, four paragraph. Let's see how that went. [بسم الله Arabic]	He is counting the paragraphs on the paper	
	1:01	Rater	Let's start with number 1, she is a Saudi lady from C1B. "in (trending) in many societies xxx in some paid work" om, uh. Right, so that is the introduction. " xx" Right, um. I am not sure about this one, " it's trending in many societies". so I just underline that one. I can check that, ok. "in some paid work xxx". Right "in recent years, it has been increasing argument whether children engagement in such works can possibly x a success" that is a good one. That is a complicated.	He is reading quietly	He draws a line under one sentence on the paper sheet. He keeps rereading the sentence more than two times
	2:03	Rater	"In recent years it has been increasing argument whether children engagement in such works can possibly (be hailed) a success, that would come to fruition". Fruition, I am not sure about this word. "in regard to helping them build their personalities or not" . Well it is a bit hard to understand what th -. "in recent years it has been increasing argument whether children engagement in such works can possibly be held the success". That is fine until here, "that would come to fruition in regard to helping them build their personalities or not". I'll have to check that word first. "in fact this matter is quite controversial there is a fine line", this is really good expression, "there is a fine line between advantages and dis -- of this issue" . that's fine,	He is rereading the first sentence again.	He seems not sure about the first sentence.
	3:01		Generally, a nice introduction although the sentence is a bit vague there in the middle. So I'll have to check what "fruition" mean. May be a spelling mistake. I have to check that in a minute.	He wrote some comments on the paper	
	3:22		"On one hand", on the one hand, there should be definite article here. "in one hand it could be argued that children labour could possibly be a hindrance	He is reading the text, the	

			toward their educational physical and psychological growth". Well that is a very excellent a topic sentence for that paragraph. This is the topic sentence, well done.	second paragraph	
	3:50		"Spending so much time, and ex ex excreting". Um, ah, probably she means exerting. I would say the wrong word. It is not excreting. "Exerting a great effort at work", for the moment I'll go back to the word "fruition" and say QUESTION mark, that word, I'll put a question mark, so that is mean it is not clear for me what word do you mean?.	The rater spells the word slowly.	On the paper he writes some abbreviations and symbols for the wrong item.
	4:26		"So spending so much time and exerting great effort at work can take it to on working child who supposed to study and perform educational skills which are learnt at school and which a child is expected to master". Well very good sentence, compound sentences written successfully. Um, xxx …. Xxxx, well excellent sentences. "xxx Some jobs require a great deal of physical effort which is a child is expected to lack", yes, "to ke—a great number of working children might suffer xxx as shown in some recent health studies". She is so intelligent, she know how to write complex sentences and she supporting her ideas. "More over there is an increasing concern that children have not obtained the adequate experience that allows them to go out and face life alone without the help and advice of the grownups". Very good ones, very good sentence. "due to that a working child might be possibly be a victim of abuse and exploitation and that in its turn can (its) great have negative impact on child …."'.		
	6:15		I would say this is a very nice and well written paragraph. So I'm writing this here. So that's the paragraph against child work.	He wrote his comments on the sheet.	
	6:46		Right now my expectation is the second paragraph should be talking about the good side of it. "on the other hand". Yes this is the good connection between the two		

			paragraphs; on one hand, on the other hand, well done. "on the other hand, it can be debated that recent studies highlighted that working at early age can provide individual and certain practical and fundamental life skills". Umm, uhah, right ….well it—the meaning is clear but umm the grammar of the sentence is not ok.		
	7:23		"On the other hand it can debated that recent studies". So that "recent studies have highlighted the fact that working at early age can provide ..". right "it can be debated that recent studies have highlighted the fact". Well the expression itself "can be debated that recent studies" it does not work, because she is not debating the studies actually. She might say on the other hand recent studies have highlighted, so I would say delete underline that one. And say delete, yes.	He is reading the first sentence of the second paragraph again. He crosses out the phrase and writes delete on the sheet	He seems accurate in his evaluation. As he go back many times to the sentences and reread them again and again.
	8:05		"It is at the heart of childrens normal growth to fell independent and be able to make choices on their own at early age since that contribute to improving their ability to be decisive in terms of their future careers". Well fantastic, fantastic sentences. A good ones. This is not C1B. it should be CC2. "it is also vital for children to understand the value of money gained after hard work which they perform themselves. Furthermore, having to deal with the pressure that is usually found in the working atmospheres and with different types of personalities". Well, "having to deal um … xx could possibly teach them to be responsible for their actions and granted them a great deal of communication and time man—skills".		'Chidrens' I am not sure whether he misspell the word or its wrong spelling and he ignore it.
	9:14		Om, well, she is genius. A well written paragraph. Excellent.	He wrote on the paper his comments.	
	9:24		"To sum up, children can get engaged in light work that suits their natural growth". Um, Ahha, that she looks like that she jumped to the conclusion straight	He is reading.	He wrote *I expected a paragraph here where*

			away. What I expected is a paragraph here. She mentioned a paragraph against child work, and the other one with child work. So she should add one more paragraph where she can identify where she is. Whether she is with or against. Right it sounds that she is writing that in the conclusion but that is too short actually. That saying "to sum up, children can get engaged in light work that suits their natural growth during holidays under the condition of being instructed and taking care of". That is fine, that is the solution for her. "the experience a working child might obtain can go hand in hand with the supervision and instruction that parents can offer toward the co x of a better future for the child".	He writes his comments on the paper.	*you show what your opinion is.*
	10:55		That is an excellent paragraph actually. It comes as a conclusion. But I would say, this paragraph should be in the body of the essay not in the conclusion. And she can repeat these words in other terms. So, I would say	He wrote the comments on the paper sheet	He wrote: *your opinion is here but this should have appeared in a body paragraph before the conclusion.*
	11:45		That's fine. It is a very nice essay. I think if she writes an essay like this in IELTS she will take something like probably 7.5, definitely 7.5. And if she provided another paragraph before the conclusion, she would get 8, perhaps. The only thing that I am not sure about "it is trending in many societies". I am not sure of that use, I have never heard of it in my entire life actually. So I will put question mark there.	He turns the page. And revaluate the beginning of the essay.	
	12:25		The other thing is, "can possibly, be held a success that co—to fruition". "Fruition", I might check that word in dictionary, I have never heard of it before, I suspect that. "fruition" it's sound English to me, but "fruition"… let us see : <u>the point in which a plan or project is realised, the realisation of a plan or a project.</u> Um um, "fruition", she is using very academic vocabulary.	He checked the word with his cell phone. He wrote in his cell phone: *fruition meaning.*	He has his cell phone.

	13:24		"in recent years, it has been increasing argument whether children engagement in such work can possibly be held success that would come to fruition", well I would say that is fine.	He reads the sentence again	
	13:47		I will check the word trend, … "it is trending in many societies" … . let's see one of the uses, I have seen it as a noun, I have never seen it as a verb. I know it is a verb, but I cannot think of an example I have seen in my life, … "so it is trending in many societies, even if that's correct, but it is kind of odd in English to say "it is trending". So, yes, to extend, inc-- x x x, um …, (the gender gap is) tending down. Yes I would accept this, right, … it is very common to have a "trend" as a noun actually, omm, to trend, to take, ... to have general tendency. I would have to accept this actually.	He is checking his cell phone dictionary. He keeps looking in the same sentence and read it again and again.	
	15:22		Xxxx ….. right …		
	15:33	Researcher	Think aloud please.		
	15:37	Rater	So, I would say that, it is ok. Although it is not frequent in English to say this. "it is trending in many societies".		
	15:52		Well, I'll delete that question mark then. And the question mark here,	He write his comments on the paper sheet.	
	16:02		Well, that is very ingenious … she can even use expression um that are rarely used in a very correct way that even myself. I am not sure about. Yes that is fine		
	16:24		So I'll write some general comment. This is a very well written essay. Right, your paragraphs are very structured and organised. Um, well, I'll say, I think you needed, because she is so clever, I think you needed to have one more body paragraph before your conclusion where you show your exact attitude or opinion about child work.	He turned the page over to the reverse side.	He wrote his overall comments as points on the back of the sheet.

	18:06		I have to admit then as well, your English is excellent. Thank you. I will write here, feedback on the back.	He reversed the paper again and wrote at the top of the paper.	
	18:45		Usually, these essays are full of grammar mistakes, but this one is so clean, well finished from the first one.		
	19:00	Researcher	How much you give her?		
	19:02	Rater	I am not giving them marks actually, but if I am marking this for IELTS, I would give it 7.5, definitely 7.5. yes.		

Appendix K: Retrospective interview transcription

T3 Zain. Script N5.

16:00	Researcher	Now, I want to ask you please about this paper. Now, this paper what grade would you give it?
	Rater	In IELTS I gave it 4.5 to be honest. If it is IELTS, I am always saying IELTS because most of the students on that module or on the ELLP, they're looking for IELTS and they keep asking about that, 'how much I will get in something like that'? So I would say in IELTS this not take more than 5. Well probably if I say 5 that's would be generous. And mainly because of Grammatical problems.
	Researcher	Now, this paper is value 4.5... what makes this paper strong? And what makes it weak?
	Rater	Well, in terms of ideas, the weak part of the essay is that at first he's saying I want to mention two main reasons why children should not work but the first paragraph he is saying children have to work sometimes but I am against this. Well he is not really mentioning a reason he just saying I am against work. So, he did not offer any reason but yes in the second part he is talking about why children should not work because ommm they might have social problems and behaviour and might you know resort to bad habits like theft and drugs and alcohol and something like this. So, I would say information is mainly concentrated in the second paragraph not the first one. So, the first one is kind, is valueless I would say it is not valuable actually. This is what I think the main danger is. In terms of GRAMMAR that is the main thing, grammar is so horrible I would say for a person in C1B (this is the advanced level in ELLP). C1B is kind of lower advanced and I would not put this more than probably in intermediate or upper intermediate perhaps.
	Researcher	So, you think his proficiency is below the level of the class he has been placed in?
	Rater	Absolutely, well I keep saying this student should not be in C1B. CIB supposed to be, oh sorry C1B is kind of advanced and C1A is lower advanced.C1B is advanced actually. And then you get C2 which is upper advanced. So this is ommm, he shouldn't be there actually.
	Researcher	Ok what makes it strong?
	Rater	Well, I will say the introduction is ok, the conclusion is ok. The ideas in second paragraph so far ok but the problem is he's got the ideas but he's not got the language to express the ideas. So the main problem is with the language itself but obviously he's got wonderful ideas in the second paragraph

		but the main thing is the language. His language is not helping actually to write what he wants to write and he looks to use the passive very often haphazardly without knowing whether he should using the passive or active. So he thinks he write complicated language by using 'have been' instead of using 'have' but this is wrong of course. And the danger is it looks like he has not got any understanding of the passive because he got 4/5 frequent mistakes. So it is not like one mistake happening once if it happens once it is fine, it might be chance or just you know he is a bit sleepy or dizzy or whatever but he got a few mistakes and these mistakes are frequent then he's got, I would not say these are mistakes, these are errors. Because mistake, it means only once but these are errors. This means he needs to get rid of them. Again he is using 'so that', 'so that' at the beginning of every sentence and you can't start a sentence with 'so that'. So he writes phrases actually rather than sentences. I would say in this essay the grammar is the main problem but in term of ideas it's fairly acceptable I would say.
	Researcher	From general perspective, what makes a good piece of writing from your point of view?
	Rater	Well, for this level we expect the marks to be deducted because lack of details, not mentioning examples, disorganization, things like these. We do not expect marks in this stage with C1B to be deducted because of bad grammar or horrible grammar like this actually. So, I would say, well a perfect essay is an essay with wonderful ideas, Organized paragraph and good grammar actually. But this is an essay, well, it has a good structure that's what I am saying, well structured essay in the sense it has introduction, conclusion and two paragraphs although one of them is slightly off point. Yeah, it varies actually from one essay to another. For this essay, the perfection of this essay would be perfecting with the grammar otherwise it is ok. You know these kinds of mistakes that somebody says 'I want to mention two main reasons' and then he mentions only one reason let's say it is very common with those students, they forget what they said in the introduction but the problem here is grammar itself. Well personally I think grammar is the most important thing, yes. Because even if you have wonderful ideas but your language does not help you, well what kind of impression you get, is like somebody has got wonderful ideas but his language is rubbish and in the end in IELTS or any other test you are evaluating language in the

		first place rather than ideas so even if you disagree with these ideas or they are not very well supported but the language is perfect so well this person can write. Of course, ideas can be developed more but the language itself rather than the content, yes. I would say a person who is B1B or about 24 years old can generally manage to write coherent ideas so they know how to reason things. They know how to structure things but if their language is bad, this is where they get lower marks in IELTS, and that's why he might go to IELTS and think well I get 7. NO this is not 7. This is 5 maximum 5. This is perhaps 4.5 in IELTS. YES. He has got some good sentences in the end and the beginning and that is why I am saying cannot be more than 5. If you look at the essay again he's got bad sentences in terms of grammar.
	Researcher	This is interesting, when you approach the essay what are the criteria that you look for in your rating first?
	Rater	Well, in general terms I am looking for the content first and then grammar, language. When I say language, I mean grammar and that includes vocabulary because for me vocabulary is part of grammar. If you know that this is an adjective not a noun then you behave accordingly. So, I divide the feedback in terms of two things: grammar and very often my comments are very local actually in the sense I underline the mistake or circle the mistake or delete the mistaken part, and second part is content. And for content, sometimes I write something next to the paragraph but very often I write feedback on the content at the end of the essay. So general feedback. Because you cannot comment on every line actually so you comment on the content very general so I keep it to the end. But in terms of grammar yes, I deal with these very quickly and locally in localized fashion. When I say grammar, I mean spelling is part of it, vocabulary is part of it. So, when the person uses a wrong word I know it is not grammar really, but still I will say this is language. We can say language and content but let me say not grammar, let's say language and content and very often the main problem for students is language itself rather than content. Even those who are in B1 or B2 they still can write wonderful ideas but the problem is they have not got the tools to express themselves. They have not got the good language to express themselves and this is where they get bad marks. If you say to them talk to me about this problem in your language they express what they know in a very wonderful way and they've got wonderful ideas perhaps better than mine actually. But the problem when they want

		to write this is where they are making mistakes and I think most of them lose marks because of bad grammar. Yeas,
	Researcher	From this paper I can see some, from your correction, you corrected wrong items, and some you just write question mark?
	Rater	Yes, question mark would be in two cases. One of them where I put question mark is when the structure is so horrible that even if you want to correct it you have to think about it too much. So, in one example he is saying: "Even more some few" "more some few" then you say what 'even a few', 'even some' or 'even more', you know I say question mark pay attention to that we cannot say "even more some few" get rid of this. Yes, in cases where the error is so striking then you can't do anything about it. I write question mark HOW STUPID IS THIS, to be honest. That is what I mean by two question marks not one. So, when I have two question marks well I am not saying this, I am not putting this on correction codes sheet but they know what I mean by two question marks, it means, this is very stupid in this level of English C1B let's say. The second time when I put question mark, like here: when he saying: "are homed" and I can't sound what he wanted to say. Is it a spelling mistake; is it another word you wanted to say? Or is it like he is not aware of the use of the word 'homed' ? or is it he's writing and thinking of something else? I am not sure what he wanted to say by "are homed". So that is why I put a question mark. It means I can't understand it. One of the codes on the correction code sheet is: a question mark means I cannot understand what you mean. One question mark indicates I cannot understand; two question marks well hang on this is very stupid, done.
	Researcher	From the think aloud, I noticed that you read and you reread, and then when you finish evaluation you also go back and reread, can you please tell more about this?
	Rater	Well, sometimes it depends, when I read the sentence for the first time very often look at the content because this is one's nature when you read something you are not looking at grammar, you think this is good grammar but you are looking at the content of the sentence. Then, while you are reading the first time you get distracted by the bad grammar of the sentence and then you have to read it again for the grammar. So, for a person I would say my first reading is for ideas, just to see what he is trying to say and then the second reading is for grammar. Sometimes if I spot a very very clear grammatical mistake I've stopped immediately without continuing the sentence. So I stopped there, I correct the

		mistake and I read again. And sometimes if I read the whole sentence without a very clear grammatical mistake I have to read it again to see where is that mistake coming from. The sentence sounds redundant or not clear or grammatically not ok or NOT ENGLISH sometimes. Well, some of the sentences even those I passed actually, well look at this sentence for example: "the only long term solution to cope with the problems of children is about economical situation", well grammatically is fine actually but this part is about economic situation it does not make sense actually. So, any reader, any native speaker who reads this sentence would say well this is not English. Although the meaning is ok but it sounds not English but still I pass it because the student got some other mistakes. The sentence is ok and it depends from one person to another. If I am marking Arwa (the Saudi lady's essay) and I found a sentence like this I would have a comment on this, I would say: it does not sound English. Because that is one of the very very few mistakes like two let's say. So if the student is very strong and got a sentence like this then I would highlight it and say this is does not sound like English. But because there are so many basic problems in this essay this sentence stands as a good one actually. That's why you cannot comment on every word otherwise it will be a mass of scribbles.
	Researcher	So, why do you not take the first impression and immediately evaluate?
	Rater	Well, if I am doing that for evaluation that is true. Now if I am marking this just to see what is the level of the student, I wouldn't give that feedback actually. See if the sheet stays with me you would not find all these grammar mistakes. I will just read it without anything and give it a mark straight away. This is good and I might underline things because at the end of the marking you might go back and say how many, for example I might underline grammatical mistakes and I might circle ideas or mistakes with ideas or content and then at the end and see the percentage of things. But for this one because it is intended to be for feedback I give feed back and write loads of things. If I am marking this for assessment and the students cannot get the feedback I would not write any comments. Yes comments for myself just to see what the mark is. The student has to learn from the feedback that's why I am keen to highlight everything.
	Researcher	So, for the feedback you read and reread and for the assessment you just read it once?
	Rater	Yes I will read it once. I might reread the sentence to understand it but you know if you are giving the mark for

		assessment even if you are looking at IELTS criteria for assessment it is not counting things like one two three if you got five mistakes you got this. It is a very general description that this student for example is able to express himself clearly with a good range of grammar and things like this. The second is this person is not able to express him or herself clearly. You got vague general criteria rather than specific things. That is why it depends with experience sometimes, if you read it once, you say well this is 5. Even here actually, when you get feedback, you get distracted by the feedback itself. Because you are commenting on so many things and the end you have to look at again and consider things generally that's why, you might have noticed when I finish the whole thing I started looking again the whole sheet in terms of paragraphs because I got distracted by the many grammar mistakes I gave. But if I am reading this for assessment I will just read it once and then perhaps I give it a mark. I still look at it again and read for example for 3 times, even if it is for assessment mark. But it is very lightly. For assessment you have to read it sometimes again, but reading would be different actually. If I am giving feedback on assessment, it would be holistic rather than analytic.
35:15	Researcher	When you correct the samples from the students to rate them, usually when do you prefer to rate them and why? How and what is your approach in rating?
		If they give me the sheets on Monday, it depends on the person. Usually I am very busy so the only time I have is one or two days before the next class. So that class I am teaching only on Monday so if I will give them the sheets on Monday I will correct them on Saturday or Sunday.
	Researcher	Why? Do you think if you correct them earlier your rating will be affected?
	Rater	You know I have other things to do. Well my rating will be affected if I am rating these sporadically. It means if I mark lets say two sheets on Wednesday and four sheets on Thursday and perhaps one or two on Sunday because if you.. mark things early you have more time to look at details and think more deeply but if you marking let's say three of four copies Sunday before Monday and have not got time, you might do things very quickly and it is not fair. That's why I keep the whole thing until the last minute. So I do them all on Sunday. Everyone gets a fair crack of the whip.
	Researcher	So you prefer to correct them all at the same time.
		Well that's my preference. Sometimes if you have two or three left, I think it is not problem actually. I might do

		something on Saturday and continue on Sunday. I do my best to maintain the same amount of feedback and the same amount of comments. I remember once I, it happened to me for some reason I forgot to mention general comments at the end of the sheet of the essay of one of the students and then he said to me: sorry why you do not mention something for me in the end because he looked at his friend's sheets and he found some general comments; he said I've got feedback here but you did not mention anything at the end. Although, I am sure he does not read what's there, but he wanted something. Yes generally I try to be fair. And give the same amount of feedback. To be honest the feedback varies between one essay to another depending on how good the essay is. If the essay is bad, you know feedback might be more general. If the essay is very very good, the feedback might be about specific things. It is like driving, if you're training somebody to drive and they cannot drive on the left lane on the road then you cannot teach them anything more - it is a basic thing that they cannot master but if they are very good drivers but at one point they neglect looking at the mirror let's say, then you tell them look at the mirror but if the driver does not know how to switch on the engine you cannot comment on this. Do you see what I mean? The better your essay is, the more specific feedback. The worse it is, the more general the feedback is.
39:25	Researcher	How do you differentiate between the levels of students? Do you give them the same rating, the same feedback? This level and lower level. Do you focus on the same traits?
39:32	Rater	Generally, in lower levels like B2 B1 ,they tend to have more grammar mistakes and then most of the feedback is on grammar actually. Because I think in this stage this is basic actually you want good grammar. For me if you do not put things in a hierarchy, grammar for me is the base of pyramid I would say, then content then ideas are higher. With lower levels like B1,2 I comment on grammar actually then the content. Because you want them to master basic structure first but when we go to C1, C2 you assume generally that their grammar is better but their ideas perhaps have problems. So more grammar with lower levels and more content related comments with high levels. Generally, the sad reality sometimes there are some students who are in C1B they should be in B1 or B2 and sometimes in B1 students who should be in C1B.

Appendix L: The final taxonomy of the Codes and their definition

Code name	Definition
I. Self-monitoring focus	
A. Judgment strategies	
a. Summarise, distinguish or tally judgments collectively	Finalise evaluation by bringing together evaluations of detailed features of the essay which have already been mentioned
b. Main cause for given score	Weight certain criteria as more relevant than others to the final overall rating
c. Ignore the error	Identify an error but seem to ignore it
d. Define or revise own criteria	Name or change criteria which are to be relied on in the evaluation/rating
e. Decide on macro strategies for reading and rating	Explicitly state what strategies / processes are going to be used, e.g. read first and then evaluate
f. Consider own personal response or biases	Voice a response to the text that is not based on linguistic or other objective criteria
g. Compare with other compositions or students	Compare text with those of other students, or the same writer on other occasions
h. Articulate or revise scoring	Give or change a numerical mark for an aspect of the essay or the essay as a whole
i. Articulate general impression	Voice an overall evaluative impression of the text, including apparently, all aspects together (language, content etc.) and given in words rather than numbers/letters
j. Interpretation strategies	
a. Teacher attitude	
01. Refer to teaching	Connect to instruction which the teacher/rater knows the writer has received
02. Teacher's expectation	Refer to what the rater expected in terms of content or organisation, language etc.
03. Provision for feedback	Say and/or write on the script something apparently directed to the writer of the text
04. Emotional reaction	Voice an effective response to the essay in general
05. Other comment	Give any other comment that is not related to evaluating the ideas, organisation or language of the text and not covered by any other IB category
b. Reflection on student's background	Refer to information known about the particular student
c. Identify student's name	Mention the student's name
d. Referring personal situation of the writer	Imagine the situation of the writer, as a way of understanding the text

e. Reading behaviour	
01. Read or interpret task prompt	Read or interpret the essay title provided and the task description
02. Read part of the text	Read sentences, paragraphs etc. interspersed with comments of any sort
03. Reread part	Read words, phrases, sentences etc. which have been read before
04. Read whole text	Read the whole text without interruption
05. Scan whole composition	Show evidence of gaining an overall impression of features of the text as a whole (e.g. length, paragraphing) without actually reading it fully
II. Rhetorical and ideational focus	
A. Judgment strategies	
a. Task fulfilment and requirement	
01. Assess relevance	Evaluate how far the content is appropriate to what the task prompt requires
02. Assess task completion	Evaluate how far a complete response is provided to what the task prompt requires
b. Relevance, quality, appropriacy of argument	Evaluate how far the argumentation and reasoning is appropriate to what the task prompt requires
c. Identify redundancy	Identify where ideas are unnecessarily repeated
d. Assess organization	
01. Paragraphing	Evaluate appropriacy of division of ideas into paragraphs
02. Main body	Evaluate appropriacy of what is in the main body of the essay
03. Introduction	Evaluate appropriacy of what is in the introduction
04. Conclusion	Evaluate appropriacy of what is in the conclusion
05. Cohesion & coherence	Evaluate appropriacy of use of linking words and connection of ideas
06. Assess reasoning, logic, or topic development	Evaluate appropriacy of the reasoning (e.g. for and against an issue), exemplification, etc.
07. Overall organization	Evaluate organisation in an undifferentiated way
B. Interpretation strategies	
e. Personal reaction to ideas	Respond communicatively and personally to the content of what is said
f. Interpret ambiguous or unclear ideas	Suggest what meaning was probably intended, where it is unclear
g. Discern rhetorical structure	Suggest what structure the text was probably intended to have, where it is unclear
III. Language focus	
A. Judgment strategies	

a. Vocabulary	
01. Choice, range, appropriacy	Evaluate the richness of the vocabulary used, and how far it is used suitably in terms of meaning and collocation
02. Errors	Identify vocabulary errors
03. General comments	Evaluate vocabulary use in general
b. Mechanics	
01. Spelling	Evaluate spelling of words
02. Punctuation	Evaluate use of punctuation marks
03. Capitalisation	Evaluate use of capital letters
c. Language fluency	Evaluate the flow of the text, including use of sentences of an appropriate length and complexity, well connected both grammatically and in thought/content, so the reader can follow the thread effortlessly
d. Consider syntax or morphology	Evaluate word order, inflections and other affixes, etc.
e. Consider error gravity or severity	Evaluate how serious an error is
f. Consider error frequency	Note frequency of a language error, either impressionistically or by counting instances
g. Clarity, confusing, comprehensibility	Comment on something hard to understand, whether because of wording, handwriting etc.
h. Assess quantity of written production	Evaluate the amount of words or ideas etc. included
i. Rate language overall	Give an overall evaluation of the accuracy and use of language
B. Interpretation strategies	
a. Observe layout	Comment on line spacing, pagination, etc.
b. Interpret phrases	Suggest what form an unclear expression was probably intended to have
c. Classify errors into types	
01. Word order, wrong word	Identify an error as involving syntax or word order
02. Verb form	Identify an error as involving verb inflection, tense etc.
03. Sing. noun, pronoun	Identify an error as involving noun or pronoun inflection, number, or countability etc.
04. Preposition	Identify an error as involving a preposition choice
05. Plurals	Identify an error as involving plurality
06. Articles	Identify an error as involving an article form or use

Appendix M: List of codes with examples

Code name	Examples
I. Self-monitoring focus	
A. Judgment strategies	
a. Summarise, distinguish or tally judgments collectively	So again, out of ten for content I would say, because of the fact she's got the structure – the introduction, the main body and the conclusion, and two good paragraphs – I would give that a seven out of ten. So, a six out of ten for grammar and vocabulary, a seven out of ten for the content, which means a thirteen out of twenty.
b. Main cause for given score	But there's no topic sentence, so I'm going to mark that down again. It's below what I'd hoped for.
c. Ignore the error	"some information or write scientific reports". Why "scientific"? I'll let that go.
d. Define or revise own criteria	In general terms, I am looking for the content first and then grammar, language.
e. Decide on macro strategies for reading and rating	OK. I'm just going to read that again because I need to focus on this question, because it's important that the students answer that.
f. Consider own personal response or biases	Well. Because sometimes with some students, in fact who may be consistently I would give 5 to, or I would imagine they would get 5 on IELTS, they do the IELTS, they come back and they say "oh! I got 6,5" oh! So, may be marking isn't exactly what IELTS would be. I prefer to go little bit lower. He may in fact be 6.5.
g. Compare with other compositions or students	Um, compared with the last one it's not nearly as, um, what should I say? It's not bad.
h. Articulate or revise scoring	So, five for number one.
i. Articulate general impression	Erm, it's not bad but it's a bit weak, I think. It's a little bit weak.
B. Interpretation strategies	
a. Teacher attitude	

01. Refer to teaching	"...younger people tend to try following cutting edge developments more than older" – "older *people*". OK, that's quite good. That's the idea we discussed in class.
02. Teacher's expectation	I was expecting one paragraph to say that they were similar and one paragraph to say they were different, but that's a bit mixed up so there's an organisational problem here.
03. Provision for feedback	Luckily, he goes straight back and he gets back on task. So, when it comes to feedback, I'll suggest to him "does this sentence really need to be here".
04. Emotional reaction	That's not right. Really disappointed in this from John.
05. Other comment	I will check the word *trend*, … "it is trending in many societies" …. let's see one of the uses, I have seen it as a noun, I have never seen it as a verb. I know it is a verb, but I cannot think of an example I have seen in my life.
b. Reflection on student's background	I think this student is Chinese and it's a common mistake which is, I think, due to transference – mistranslation from his language.
c. Identify student's name	OK, let's look at Amjad.
d. Referring to personal situation of the writer	You know the other thing about this guy? He's a really popular student. Like I say, when he first asked questions people were rolling their eyes and tutting; now when he asks questions you can see everybody will be thinking "that's a good question
e. Reading behaviour	
01. Read or interpret task prompt	"The question is what *are* the differences between two generations" – hmm, OK "between the two generations' methods".
02. Read part of the text	"However older people tend to use" – OK, good, we need a comma after "However" – "older people tend to use applications for their basic life". OK.
03. Reread part of the text	Young people prefer to use modern technology for chatting in social networks" – there's no topic sentence – "playing video games but sometimes..." "Young people prefer to use

	modern technology for chatting in social networks, playing video games but sometimes for gets" – "to get"
04. Read whole text	When I mark all these papers I read them all first, without any correction. So I've actually read all of these papers very quickly, just to see if I could read them, just to see if they make sense.
05. Scan whole text	I've asked for at least 150 words, and just going through it I can see that it's around about that amount.
II. Rhetorical and ideational focus	
A. Judgment strategies	
a. Task fulfilment and requirement	
01. Assess relevance	OK, good news is it's on topic and it's that bit has been done well, actually.
02. Assess task completion	They haven't finished the essay.
b. Relevance, quality, appropriacy of argument	Ok, that's quite a good paragraph, actually, although there are few language mistakes. He's expressed his ideas well and I think that's quite good argumentation, so that's good.
c. Identify redundancy	A little bit of repetition there because, looking at this paragraph, she's already said "it has lots of houses and many trees" and she's now saying "there was no factory and the seaside has many trees"
d. Assess organization	
01. Paragraphing	I was expecting one paragraph to say that they were similar and one paragraph to say they were different, but that's a bit mixed up so there's an organisational problem here.
02. Main body	The main body is more about that. So, it does have organisation.
03. Introduction	Generally, a good introduction.
04. Conclusion	Does the conclusion match? Yes, the conclusion matches the introduction.

05. Cohesion & coherence	And part of that is the organisation. Erm, some of the linking is OK, in that we've got "for example", "for instance".
06. Assess reasoning, logic, or topic development	- He's sort of following a sequence; he's organising his ideas well. - He hasn't developed the argument as much as I would have liked, really. And I don't know if that works.
07. Overall organization	Organisation is as expected.
B. Interpretation strategies	
a. Personal reaction to ideas	- Well, that is a bit silly. She still believes in Darwin. - "Even if they have time they spend it doing gardening" – ha ha – "or walking because they think spending more time using technology will affect their health." I'm not sure that's a very good picture of older people. I'm not sure that's even true, but that doesn't matter
b. Interpret ambiguous or unclear ideas	"outweigh the advantages whereas sometimes fast food lead to death as we know there is not anything more important than human soul". Ok, this is very messy. Instead of "whereas", I think he needs to write 'because'. "the younger generation tend to have plenty of time and generally stay late at night". Stay late at night? Stay in a hotel? What? come on. 'stay up late at night'. I guess. Yep.
c. Discern rhetorical structure	Next paragraph – I think it's a new paragraph.
III. Language focus	
A. Judgment strategies	
a. Vocabulary	
01. Choice, range, appropriacy	"Widespread" at the beginning – "widespread" is good. "Vast majority" is good. "Send emails" is the right collocation.

02. Errors	"spend their time to make friendships with each other" – "to make *friends* with each other". Nearly.
03. General comments	Vocabulary? Well, there's some good vocabulary in there.
b. Mechanics	
01. Spelling	"*of* diferent ages". Spelling of "different" is wrong
02. Punctuation	"For example," – comma is good.
03. Capitalisation	"in terms of health" should be capital "I".
c. Language fluency	Ok, that's quite good. He's shown quite good accuracy, accurate use of language mostly, and vocabulary and he's given a couple of examples. Yes, I am not sure he's given several reasons- he's given two, but that's ok. I mean, it's a short essay. That's good. I like it.
d. Consider syntax or morphology	The student used relative clause with 'which'.
e. Consider error gravity or severity	...so that's quite a significant spelling mistake.
f. Consider error frequency	"economically if person", again, -"if people", repeat the same message- "if people stay long time"-
g. Clarity, confusing, comprehensibility	"Such as weakening eyesight effect on" – can't read this – "effect" or "affect on span"? Lifespan, I wonder?
h. Assess quantity of written production	So, that's quite good. He hasn't covered a lot of points; ut again, the length of this essay doesn't really allow the students to write a lot. But it's quite a good essay.
i. Rate language overall	He's used the language very well, minor mistakes.
B. Interpretation strategies	
a. Observe layout	So here I see I've got from Douar one piece of paper on two sides, written double spaced.
b. Interpret or edit phrases	OK – "which are *suitable*", I think he means, "which are suitable for them".
c. Classify errors into types	

d. Word order, wrong word	"tend to do not that" – "tend not to", wrong word order.
e. Verb form	He's got confused, muddled up tenses, he's used... He started using past, but he's used the wrong tense – the continuous, not the simple.
f. Sing. noun, pronoun	... again, should be uncountable.
g. Preposition	"it lets young people easily get information with meeting their need". Yes, so wrong preposition.
h. Plurals	"technology makes person" – again, must use the plural in general. I'm going to make a note of that – use plural for general.
i. Articles	Missing his article 'the' before "library".