IELTS™

An investigation of examiner rating of coherence and cohesion in the IELTS Academic Writing Task 2

Authors
Fiona Cotton
University of New South Wales

Kate Wilson
University of Canberra

Grant awarded Round 14, 2008

This study takes an in-depth look at the assessment of coherence and cohesion (CC) in the IELTS Academic Writing Task 2. It investigates the level of difficulty examiners experience, the features they look for, and the extent to which their marking of CC differs from their marking of other criteria. The impact of examiner qualifications, experience and training materials on assessment reliability is also examined.

Click here to read the Introduction to this volume which includes an appraisal of this research, its context and impact.

ABSTRACT

The study investigated whether examiners find the marking of coherence and cohesion (CC) in the IELTS Academic Writing Task 2 more difficult than the marking of the other criteria; what features of CC examiners are looking for in marking Academic Writing Task 2; the extent to which they differ in their marking of CC compared to their marking of the other criteria; whether qualifications and experience had an impact on assessment reliability; and how much current examiner training materials clarify understandings of CC.

The study involved think-aloud protocols and follow-up interviews with 12 examiners marking a set of 10 scripts, and a quantitative study with 55 examiners marking 12 scripts and completing a follow-up questionnaire.

The quantitative data revealed that examiner reliability was within the acceptable range for all four criteria. The marking of CC was slightly less reliable than the marking of Grammatical Range and Accuracy and Lexical Resource, but not significantly different to Task Response. No significant effects could be found for examiners' qualifications or experience, which suggests that the training is effective. The findings showed that examiners found the marking of CC more difficult than the other criteria.

Examiners were conscientious in applying the band descriptors and used the terminology of the descriptors for CC most of the time. They also introduced other terms not explicitly used in the CC descriptors, such as 'flow', 'structure' and 'linking words', as well as the terms, 'essay', 'introduction' 'conclusion' and 'topic sentence'. The introduction of terms such as these, together with variation in the degree to which examiners focused on particular features of CC, has implications for the construct validity of the test.

Suggestions for improving the construct validity include: possible fine tuning of the CC band descriptors; clarification of the expected rhetorical genre; further linguistic research to provide detailed analysis of CC in sample texts; and refinements to the training materials, including a glossary of key terms and sample scripts showing all cohesive ties.

AUTHOR BIODATA

FIONA COTTON

Fiona Cotton (BA, Dip Ed, RSA Cert TESOL, M App Ling) was until recently Senior Lecturer in English Communication at the University of New South Wales at the Australian Defence Force Academy. She is founder of the Academic Language and Learning (ALL) Unit and coordinated the program from 2006–2009, for which she won a Learning and Teaching Award in 2006. Before being employed in her current position, she taught ESL for many years in Asia and Australia. Her current teaching and research interests include academic writing and literacy development in university contexts. She has been an IELTS examiner since 1994.

KATE WILSON

Kate Wilson (MAHons, Dip Ed, MEd by research, PhD) is an independent researcher and Adjunct Associate Professor of the University of Canberra. She was formerly Director of the Academic Skills Program at the University of Canberra, and Head of the School of Languages and International Education. She has extensive experience in English language teaching and research, including 10 years as an IELTS Examiner, and 20 years' experience in English for Academic Purposes (EAP) both as teacher and teacher educator. Her doctoral research, as well as her masters by research, have both concerned international students' academic literacy.

IELTS RESEARCH REPORTS, VOLUME 12, 2011

Published by: IDP: IELTS Australia and British Council
Editor: Jenny Osborne, IDP: IELTS Australia
Editorial consultant: Petronella McGovern, IDP: IELTS Australia
Editorial assistance: Judith Fairbairn, British Council
Acknowledgements: Dr Lynda Taylor, University of Cambridge ESOL Examinations

IDP: IELTS Australia Pty Limited
ABN 84 008 664 766
Level 8, 535 Bourke St
Melbourne VIC 3000, Australia
Tel +61 3 9612 4400
Email ielts.communications@idp.com
Web www.ielts.org

British Council
Bridgewater House
58 Whitworth St
Manchester, M1 6BB, United Kingdom
Tel +44 161 957 7755
Email ielts@britishcouncil.org
Web www.ielts.org

National Library of Australia, cataloguing-in-publication data, 2011 edition, IELTS Research Reports 2011 Volume 12
ISBN 978-0-9775875-8-2

CONTENTS

1 INTRODUCTION

This research investigated the assessment of coherence and cohesion (CC), the second criterion for assessing writing performance in the IELTS Academic Writing Task 2. Of the four criteria for marking IELTS writing, there is anecdotal evidence to suggest that evaluating coherence and cohesion is more subjective than for the other three criteria and depends to a significant extent on individual markers' perceptions of what features constitute a coherent and cohesive text. Additional feedback from a number of IELTS trainers indicates that examiner trainees seem to experience more difficulty evaluating CC than the other criteria (Grammatical Range and Accuracy, Task Response and Lexical Resource).

The CC criterion was introduced into the assessment of Task 2 in 2005, when a set of revised IELTS band descriptors was introduced after a long period of extensive research and consultation (Shaw and Falvey, 2008). The revisions aimed to remove examiner use of holistic marking and to strengthen the analytic quality of the assessment. They included the introduction of four, rather than three, criteria and more detailed wordings of the band descriptors to enable examiners to be more precise in their marking. Although the new descriptors were well received and considered to be a major improvement on the earlier scales, feedback from IELTS examiners in the trialling of the revised rating scale indicated that they tended to find the assessment of CC more difficult than the assessment of the other four criteria (Shaw and Falvey, 2008, p 165).

While both coherence and cohesion are essential for connectedness in text, Jones (2007) suggests that coherence tends to depend more on reader interpretation of the text and top-down processing, whereas cohesion depends on explicit linguistic elements of the actual text and involves bottom-up processing. It is possible that some examiners may pay greater attention to the identification of some of these explicit grammatical and lexical elements of cohesion than to others, and that insufficient attention may be paid to propositional coherence. As Canagarajah (2002, pp 60-61) has pointed out, a text can contain many cohesive devices but lack meaning. These observations about examiners' rating of CC suggested the need for a more comprehensive research study.

This study, therefore, sought to investigate which aspects individual markers identify within the writing scripts as contributing to their assessment of coherence and cohesion in the IELTS Academic Writing Task 2; the extent to which markers varied in the rating of CC in Task 2; and the ways in which factors such as the examiners' qualifications and experience affected their rating of this criterion.

More specifically, the study addressed the following questions with the main focus on Question 2:

1. Do examiners find the marking of CC more difficult than the marking of the other three criteria?
2. What are examiners looking for in marking CC in Task 2? What features of Task 2 texts affect their decision-making in relation to the CC band descriptors?
3. To what extent do examiners differ in their marking of coherence and cohesion in Task 2 of the Academic Writing module?
4. What effect do variables such as examiners' qualifications and experience have on their marking of coherence and cohesion?
5. To what extent do existing training materials clarify examiner perceptions of coherence and cohesion?

The results from this study are intended to provide insights to assist in the development of the examiner training materials or procedures and may also be of relevance in any future revisions of the descriptors. Such research is important at a time when IELTS is expanding globally. As Hamp-Lyons (2007, p 3) points out, the larger the group of examiners, the more difficult it can be to maintain inter-rater reliability and the greater the importance of examiner training.

2 LITERATURE REVIEW

2.1 Coherence and cohesion

Research on coherence and cohesion and their assessment falls broadly within the theoretical framework for the conceptualisation of communicative competence proposed by Canale and Swain (1980) and further developed by Canale (1983; 1984). They proposed that communicative competence includes four key areas: grammatical competence, socio-linguistic competence, strategic competence and discourse competence. Canale (1983, p 3) indicated that discourse competence, an aspect of communicative competence, referred to the means whereby a text develops unity through the use of both cohesion and coherence. He indicated that cohesion refers to the connectedness provided by structural cohesive devices such as pronouns and synonyms, while coherence refers to the way in which the relationships between different semantic meanings unify a text. Canale's definition is reflected in that of Shaw and Falvey (2008, p 42) who state that:

> *Coherence refers to the linking of ideas through logical sequencing, while cohesion refers to the varied and apposite use of cohesive devices (eg logical connectors, pronouns and conjunctions) to assist in making the conceptual and referential relationships between and within sentences clear: coherence is conceptual while cohesion is linguistic.*

These definitions suggest that while cohesion is an overt feature of text that is open to analysis, coherence is a more subtle feature which lies, at least to some extent, with the reader and his/her ability to make meaning from the text. As Hoey (1991, p 12) puts it, 'coherence is a facet of the reader's evaluation of a text' while 'cohesion is a property of the text'.

2.1.1 Coherence

While coherence is arguably more difficult to define and analyse than cohesion, thematic progression has been proposed as one way in which meaning is developed in text. Halliday, following the Prague School of Linguistics, saw text as composed of clauses, in which the theme – what the clause is about: 'the point of departure for the clause' (Halliday and Matthiessen 2004, p 64) – is developed in the rheme, which presents new information about that theme. Typically, this rheme is picked up as the theme of later clauses in the text, either in an adjacent clause or some time later in the text, contributing to the 'discourse flow' (pp 87-88). Halliday pointed out that paragraphs, and indeed whole texts, also have a thematic pattern.

Rhetorical Structure Analysis is another approach to analysing coherence, proposed by Mann and Thompson (1989).The text is analysed in terms of hierarchical relations between nuclei and satellites, each nucleus being the key proposition and the satellite being the way in which this nucleus is supported. Mann and Thompson identified 20 different ways in which the satellites relate to the nuclei, including elaboration, concession and evidence.

Another way in which propositional coherence has been investigated is through topic-based analysis. According to Watson Todd (1998), topic-based analysis involves a top-down approach and makes use of schemata theory. Content schema usually describe in hierarchical terms a series of related topics or propositions in tabular or tree diagram form. Topic-based analysis involves analysing the ways in which topics evolve and change over a stretch of text. In analysing spoken discourse, Crow (1983) identified six ways in which topics may progress. These include topic maintenance, topic shift, non-coherent topic shift, coherent topic shift, topic renewal and topic insertion. However, there are problems with topic-based analysis because of the subjectivity involved in pinning down particular topics and their relationships, and following their progression through a text.

Topic Structure Analysis (TSA) is an approach to analysing coherence building on the work of Halliday and the Prague School of Linguistics. TSA has been used to identify different categories of thematic progression, the most common being sequential progression where the rheme of one sentence becomes the theme of the next (a-b, b-c, c-d), and parallel progression where the theme of one clause becomes the theme of the next or subsequent clauses (a-b, a-c, a-d). Alternatively, in extended parallel progression, the first and the last topics of a piece of text are the same but are interrupted with some sequential progression (a-b, b-c, a-d). Studies referring to this approach include those by Connor and Farmer (1990) and Schneider and Connor (1990). While studies of thematic progression are a valuable way of analysing coherence in text, they do not, however, take account of all features of coherence.

One such aspect of coherence not addressed by TSA is the overall organisation of the text. Rhetoric studies have shown that certain text-types are characterised by particular features – including characteristic stages – which 'help people interpret and create particular texts' (Paltridge 2001, p 2). One of the most familiar genres to English teachers (and examiners) is the 'essay' with its characteristic introduction–body–conclusion structure. Connor (1990), for example, found that the single most important factor in explaining the marking of three experienced markers of 150 NS essays was the Toulmin measure of logical progression, which identifies 'claim–data–warrant'. These characteristic stages of the essay structure are deeply embedded into academic English writing curricula (see Cox and Hill 2004; Oshima and Hogue 2006, for example). However, research has shown that the essay genre is culture-specific. A study by Mickan and Slater (2003), for example, compared the writing of six non-native speakers (NNS) (including four Chinese) and six native speaker Year 11 students. It found that the native speakers (NS) used an opening paragraph to establish a position and a closing paragraph to restate their point, whereas the NNS were much less transparent in establishing a point of view. Even if they rounded off their text, the NNS generally did not present a conclusion, so that their writing appeared as a discussion rather than an answer to the question.

2.1.2 Cohesion

Analysis of cohesion must include an approach which identifies the explicit lexical and grammatical items which bind a text together. The most influential approach to cohesion to date was developed by Halliday and Hasan (1976) who identified five distinct categories: reference, substitution, ellipsis, conjunction and lexical cohesion. Reference chains are created largely by the use of personal and demonstrative pronouns, determiners and comparatives, linking elements within a text through anaphoric, and to a lesser extent cataphoric, relations. Conjunction establishes logico-semantic cohesive ties through the use of conjunctive 'markers' which 'move the text forward' (Halliday and Matthiessen 2004, p 535). Ellipsis and substitution allow for parts of a sentence to be omitted in referring to an earlier verbal or nominal element (for example: *I told you SO; I've got ONE*). Lexical cohesion is produced through the use of repetition, synonymy, meronymy and collocation. These grammatical and lexical means of creating cohesion Halliday refers to as 'cohesive devices'.

Hoey's (1991) approach to cohesion focused particularly on lexical ties in text. He suggested that text is 'organised' rather than 'structured' and that 'well-bonded' sentences have at least three ties to other sentences in a text, creating 'inter-related packages of information' (p 48). Thus, sentences together have a meaning that is greater than the sum of their parts (p 13). In addition to Halliday's categories of lexical cohesion, Hoey introduced the notion of 'cohesive breaks'. Watson Todd et al (2007) argue that if these 'cohesive breaks' are the points in which communication breaks down, then perhaps Hoey's approach might be more useful than Halliday and Hasan's in the analysis of cohesion. Hoey pointed out that 'the presence of a cohesive tie can predispose a reader to find a text coherent' (p 12). However, he warned that texts which are strongly cohesively bonded may lack coherence because of over-repetitiveness or poor logical links.

2.2 The role of the band descriptors

Various researchers have suggested that rating variance may relate to the vagueness of the descriptors in different rating scales (Watson Todd, Thienpermpool et al 2004; Watson Todd, Khongput et al 2007). As Shaw and Falvey (2008, p 12) state: 'The development of a scale and the descriptors for each scale level are of critical importance for the validity of the assessment.' Calls have been made for more research to ensure rating scales are based on sound empirical studies of sample written texts (North and Schneider 1998; Turner and Upshur 2002).

One such empirical study is that by Knoch (2007) who developed a scale for measuring coherence using a TSA approach, based on analysis of over 600 expository texts. Her scale included the following variables: direct sequential progression, indirect progression, superstructure, coherence breaks, unrelated sequential progression, parallel progression and extended progression. Eight raters, trained in the use of the new scale, were able to rate 100 university diagnostic scripts more consistently and achieved greater similarity in their rating than when using the original multi-trait scale, which included organisation, coherence, development, as well as style among its nine traits. The TSA scale allowed markers to analyse text closely by identifying thematic progression in detail, giving a more objective assessment of coherence. Nevertheless, recognising thematic links still relies to a certain extent on the perception of the reader. A further criticism of the TSA scale is that it does not address all aspects of coherence.

The vagueness of the descriptors in different rating scales also raises the question of construct validity. Weigle (2002) and McNamara (1996) observe that band descriptors embody the construct being assessed, revealing the theoretical basis from which they are developed. To ensure the construct validity of any test, these descriptors must describe as clearly as possible that which is being assessed in a way that can be readily understood by the users. This has implications for the assessment of coherence and cohesion in particular because, as Knoch (2007) suggests, difficulties in rating may be related to difficulties in operationalising these constructs.

In the case of the revised IELTS descriptors, a decision was made to favour analytic over holistic marking, as it produces a greater number of observations, reduces the possibility for impressionistic rater bias, and discourages norm-referencing (Shaw and Falvey 2008, p 37). The descriptors were revised on the basis of a number of research studies, particularly Kennedy and Thorp's analysis of a corpus of sample IELTS scripts (reported in Kennedy and Thorp, 2007) and the Common Scale for Writing studies reported in Hawkey (2001). The descriptors underwent an iterative process of trialling and redrafting by two teams of independent raters. Sample scripts were analysed against the revised descriptors, and both quantitative and qualitative validation studies undertaken (Shaw and Falvey, 2004). However, Shaw and Falvey (2004) used only 15 raters in their quantitative study, all of whom were experienced examiners and, as they point out (Shaw and Falvey 2008, p 13), ongoing validation studies are necessary to ensure confidence in the rating scales.

2.3 Examiner characteristics

Factors such as the background and experience of the examiners have also been shown to affect rater reliability (Hamp-Lyons, 1991; Milanovic, Saville and Shuhong, 1996; Wolfe, 1997). As North and Schneider (1998) have put it, 'however good descriptors are and however objectively they are scaled, they are still subject to interpretation by raters in relation to groups of learners' (p 243).

Eckes (2008, p 156) points out that raters may differ not only in the way they understand and operationalise the criteria, but also in the degree to which they comply with the scoring rubric, the degree of severity or leniency they apply, and in the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks.

Various studies have pointed to differences in examiners' style of marking. Wolfe (1997), in an analysis of 36 scorers marking narrative essays, found that the more reliable scoring was generated by examiners who were systematic in their approach, who read the essay before assigning a grade and made fewer 'jumps' between scoring categories. A second study demonstrated that more reliable and consistent raters focused at a more general level and stayed closer to the scoring rubric than less proficient raters (Wolfe, Kao et al 1998). In highly detailed analyses of think-aloud scoring processes with a limited number of subjects, DeRemer (1998) and Lumley (2002, 2005) have shed further light on the complex problem-solving processes used by examiners. DeRemer, analysing three examiners only, found that one of them attempted to match his/her response to the text and the language of the scoring rubric, while a second examiner got a quick impression and immediately assigned a grade, and the third examiner tended to consider the rubric carefully first before assigning a grade (DeRemer 1998). DeRemer characterised these approaches as general impression scoring, text-based evaluation, and rubric-based evaluation. Lumley (2002) stressed the highly complex nature of the scoring process. The four examiners in his study first gained a global, intuitive impression of the script and then justified this against the band descriptors to produce a final score. However, global, holistic impressions are generally criticised as being both less reliable and less valid (Allison 1999; O'Sullivan and Taylor 2002, cited in Shaw and Falvey 2008, p 28).

Examiner background may also be a factor affecting the reliability of marking written scripts. Eckes (2008) attempted to correlate marking style with examiners' background. In a survey-based study of 64 markers of a writing task with German as the foreign language, he asked examiners to prioritise the features of text they considered to be important in their marking. Eckes identified six rater types, of which four were dominant: the Syntax Type, the Correctness Type, the Structure Type, and the Fluency Type. He found that some examiner characteristics showed positive correlations with marking preferences. For example, older examiners were less likely to favour Fluency; and raters who spoke a number of foreign languages were more inclined to focus on Syntax, while those who did not speak many languages tended to focus on Fluency. A study by Barkaoui (2007) in Tunisia showed that, as they mark, raters used 'internal criteria' strongly influenced by their own teaching, despite extensive training. It is possible, as Eckes found, that factors such as education, teaching experience and marking experience may influence the way examiners construct understandings of CC and their approach to marking.

Further study using think-aloud protocols has been recommended as a particularly appropriate methodology for exploring subjective marking processes by several authors (such as Furneaux and Rignall 2007; Kennedy and Thorp 2007, Shaw and Falvey 2008). A number of studies of subjective marking processes have used this methodology (for example: Wolfe 1997; Brown 2000; Cumming, Kantor et al 2001; Lumley 2002; Shaw 2006; Barkaoui 2007). In particular, Milanovic, Saville and Shugong (1996, p 93) point out the relevance of such studies for improving examiner training.
As Lumley (2002) and Brown (2000) stress, verbal protocols do not by any means provide a complete account of the complexity of examiners' marking processes as examiners may only partially verbalise their thought processes, and may not even be aware of deeply internalised reactions to candidates' writing. However, they do have the potential to provide rich data about the cognition of examiners as they assess scripts and, as in the case of this study, about the features of text which are the focus of their attention.

2.4 Examiner training

The quality and type of rater training has a particular bearing on the assessment of writing performance (Weigle 1994; Wolfe 1997; Weigle 1998). To overcome variability between examiners, training is essential. A number of studies have found that both rater consistency and inter-rater reliability in writing tests can be improved through training (for example: Weigle 1994; Weigle 1998; Knoch, Read et al 2007; Schaefer 2008). According to Hamp-Lyons (2007), trainees should leave the training feeling confident, rather than confused, frustrated or, on the other hand, opinionated and over-confident. They need to develop a sense of participation in a community of practice and a common language to articulate their analysis of scripts. A survey by McDowell (2000) suggests that this is generally achieved in the IELTS training process. Forty five IELTS examiners in 12 countries responded to the survey. They were generally very positive about the training, although they were less sure about their preparation for marking Task 2 than Task 1. Many examiners would have preferred more problem scripts and there was strong agreement that they benefitted from the 'homework' scripts, which are no longer available to trainees. Schaefer (2008, p 469) suggests that the training could be enhanced by using multi-faceted Rasch analysis to produce 'assessment maps' of each rater's marking so that raters can become more aware of their bias patterns. Shaw (2002, p 17) also discusses whether consensus-style training as opposed to a top-down style of training might be more effective and calls for further research in this area.

In summary, the literature of relevance to this study raises some interesting questions about the clarity of the CC band descriptors, and the ways in which examiners interpret these descriptors, as well as the degree to which training assists examiners to implement the descriptors. This study attempts to explore some of these questions.

3 METHODOLOGY

A mixed method study was devised to include both a qualitative phase and a quantitative phase. The main focus of Phase 1, the qualitative phase involving 12 examiners, was to explore in some depth examiner perceptions of, and training in, the assessment of CC (Research Questions 1, 2 and 5). The first objective of Phase 2, the quantitative phase, was to investigate these three questions further through a survey of 55 examiners. The second objective, related to examiner reliability, was to explore through a statistical analysis Research Questions 3 and 4 – the extent to which examiners differed in their marking of coherence and cohesion against the standardised scores compared to their marking of the other criteria, and the degree to which variables such as qualifications and experience had an impact on rater scoring in this population.

Ethics clearance for the study was obtained from the University of Canberra's Committee for Ethics in Human Research and all research personnel signed confidentiality agreements. Official access was given to the IELTS examiner training materials under secure conditions. An initial overview of these training materials, together with the 'Instructions for Examiners' booklet, and the band descriptors for CC was undertaken to identify the key concepts underpinning the scoring system for IELTS CC.

3.1 Phase 1: Qualitative phase

The qualitative phase of the study used both a think-aloud protocol, recorded as examiners were in the process of marking, and a follow-up semi-guided interview.

Twelve volunteers were recruited from two testing centres in two different Australian cities: six examiners with less than two years' experience as IELTS examiners, and six with more than five years' experience. They comprised three males and nine females. Participants were paid at normal marking rates. Each examiner and both testing centre administrators signed official ethics approval forms for their participation in the project and were also bound by the normal IELTS confidentiality conditions not to divulge information about this research project. To ensure anonymity, participating examiners are referred to by their identifying initials throughout this report. To minimise the impact of prior knowledge on their think-aloud reports, participants were given only the most essential information about the purpose of the study before their participation.

Each examiner marked a set of 10 standardised Academic Task 2 scripts across all four criteria, following the normal procedures. The standardised scripts together with their bandscores were provided by Cambridge ESOL and covered a representative range of levels. All scripts addressed the same Writing Task A (see Appendix 1). The first five scripts were marked following standard IELTS marking procedures. After a few minutes break, the second five scripts were marked across all four criteria under 'think-aloud' conditions or what Shaw and Falvey (2006, p 3) refer to as *temporal – concurrent*, that is examiners talked aloud or verbalised their thoughts at the same time as they were assessing the scripts.

Examiners were asked to mark the first five scripts in the normal way so that they would be thoroughly familiar with the task and possible types of response before they marked the second five scripts using the 'think-aloud' procedure, which was unfamiliar to most participants.

Ericsson and Simon (1984) and Faerch and Kasper (1987) suggest that, in order to overcome the possible limitations of introspective research methods, several factors need to be taken into account. Given that cognition is an essentially private activity (Padron and Waxman 1988), most examiners are not familiar with verbalising cognitive processes, so the think-aloud process was carefully explained and illustrated. Examiners were encouraged to voice whatever was going through their mind as they marked the scripts, whether they were reading the script or the descriptors, or deliberating on a grade.

As participants can be self conscious about researchers 'listening in' to their internal voices, they were reassured that the project was a non-evaluative study and that all data would be de-identified and kept confidential.

Because the think-aloud procedure adopted for Phase 1 of this research study is different from the normal procedures for marking IELTS scripts, it is possible that the 12 participating examiners may have assessed scripts differently from the way they would have marked under normal conditions. Therefore, no attempt was made to assess the reliability of the Phase 1 examiners' assessments against the standardised scores. Nevertheless, think-aloud protocols offer a unique insight into examiner cognition which is not available through other means (Falvey and Shaw, 2006, p 3). To triangulate the data, follow-up interviews were also conducted and the qualitative data was further matched against the quantitative data in Phase 2 of the study.

Immediately on completion of the think-aloud recording, each examiner participated in a semi-guided interview lasting from 30 minutes to one hour.

The semi-guided interview schedule (Appendix 2) included questions to probe:

- examiners' perceptions of the different criteria
- their views on the band descriptors
- specific features of CC which affect their decision-making
- their views of the training in relation to CC.

Examiners were then asked to comment on their assessment of CC in the scripts they had just marked. Both the think-aloud protocols and the interviews were recorded.

Measures were taken to increase the validity and reliability of both the interview schedule and the think-aloud protocols, and at the same time to ensure the smooth organisation and timing of the data collection process for Phase 1. These measures included a series of discussions with a number of experienced IELTS examiners, the refinement and piloting of the interview schedule and the trialling of both the think-aloud process and the follow-up interview.

The recordings of both the think-aloud protocols and the semi-guided interviews were transcribed by a research assistant, carefully supervised by the researchers. The transcripts were extensively checked before being broken up into segments.

The analysis of the think-aloud protocols in the study, involving the segmentation and the coding of each segment of the transcripts, was derived mainly from the work of Green (1998) and Lumley (2005). The segmentation was based on units of meaning at the level of clause, although where a single idea carried over into the next meaning unit, both were included in a single segment; when examiners were reading from the script each incidence was recorded as one segment.

Segments were coded at four levels:

1. Firstly, each segment was coded to identify the examiners' general behaviour while marking. These behaviours included: managing the assessment process; reading either the script, the criteria or the question; judging the script; or interpreting the meaning intended by the writer of the script.
2. The same segments were then coded to identify each examiner's specific behaviour while making judgements during the marking process, such as evaluating the scripts or part thereof, hesitating, grading or justifying their grading decisions.
3. Segments were then coded to identify in general terms what it was examiners were referring to while making their judgements, for example, whether they were making judgements about the whole text, the application of the individual criteria (either TR, CC, LR or GRA) to the scripts, and occasionally the testees themselves.
4. Finally, only those segments referring to coherence and cohesion (CC) were analysed and coded to identify the specific features of both coherence and cohesion that examiners were assessing. Examples of these features include not only those taken from the band descriptors such as: logical organisation, progression, paragraphing, discourse markers, reference and substitution, but also other terms such as 'flow', linking words' and 'overall structure' that examiners used in their think-aloud recordings.

While the think-aloud data provided information about the cognitive processes of individual markers as they undertook the complex task of assessing all four criteria in the scripts, the focus of analysis for this study was on the sections of transcript directly related to the assessment of CC. The data was independently coded by the two researchers, and carefully cross-checked for consistency. Some segments required multiple codes while others remained ambiguous. At times, coding tended to be interpretive rather than definitive. For the purposes of this paper, therefore, we only report on those segments where examiners made explicit reference to one of the major features of CC or referred to examples of these features in their assessment (See Table 1).

	Features of CC explicitly discussed	Codes	Example segments
1	coherence	COH	Coherence, well they're trying. They're trying. (M/182)
2	meaning/message/ideas	M	You can certainly see what he's trying to say. (B/446) You can get a message there I suppose. (D/387)
3	argument	ARG	this argument is not coherent (A/18)
4	flow/fluency	FL	but it's the overall flow is OK (F/662)
5	clarity	CL	it's certainly not as clear as an 8 (L/30)
6	logic	LOG	what he's got to say is logical (K/79)
7	logical organisation	LOG ORG	on the whole it's logically organised (J/50)
8	logical progression	LOG PRO	and there's no clear progression.(L/218)
9	logical relationships/ semantic links	REL	Um, yep, they [the ideas] are - they relate to each other (E/191)
10	paragraphing	PARA	Paragraphing doesn't look as good. (D/152)
11	introduction	INTRO	OK introduction's pretty sloppy (M/212)
12	conclusion	CONCL	and the - probably not complete, incomplete conclusion is open ended (B/449)
13	cohesion	CO-HESION	Um, it's fairly high in terms of cohesion I think (S/125)
14	cohesive devices	CD	yeah, there is certainly a range of cohesive devices (L/41)
15	coordinating conjunctions	CONJ	So there's a problem with the coordinator there (K/4) He's got some idea of basic conjunctions as well as basic transition signals. (S/217)
16	discourse markers/ link words	DM	So automatically I'm drawn to the fact that the discourse markers are way off. (K/115)
17	reference	REF	Reference is OK. Um (S/345)
18	substitution	SUB	It's more the lack of substitution, um makes it seem very repetitive (K/32)

Table 1: Features of CC and their codes explicity referred to in the think-aloud data

The list of all codes including those related to examiner behaviours can be seen in Appendix 3. The extensive use of Excel facilitated the segmentation, coding and analysis of the data.

3.2 Phase 2: Quantitative phase

Fifty-five examiners were recruited from four different testing centres. They comprised 22 males and 28 females and five unidentified in the survey data. The examiners were employed under the same conditions as for the participants in Phase 1. Their biodata can be seen in Appendix 4.

Examiners marked 12 standardised Academic Task 2 scripts provided by Cambridge ESOL – six representative scripts at different levels in answer to Writing Task A, and six in answer to Writing Task B (Appendix 1). The original intention was to use the same set of 10 standardised scripts for both phases of the study. However, several examiners in Phase 1 raised questions in relation to the wording of Academic Writing Task A (see section 4.3.7). To minimise the possible effect of question type or wording on examiner marking in Phase 2, it was decided that Phase 2 examiners would mark six scripts for Task A and six scripts for an alternative Writing Task B provided by Cambridge ESOL. The scripts included every level from Band 3 to Band 8.

To counter any script order effect on examiner marking, the scripts were sorted into four groups and distributed at random to the examiners:

- Task A, Scripts 1-6, followed by Task B, Scripts 1-6
- Task A, Scripts 6-1, followed by Task B, Scripts 6-1
- Task B, Scripts 1-6, followed by Task A, Scripts 1-6
- Task B, Scripts 6-1, followed by Task A, Scripts 6-1.

Although this was an experimental study, every effort was made to ensure that data collection in both Phase 1 and 2 followed the normal conditions of marking as closely as possible to minimise the impact of the research design on our findings.

After marking the 12 scripts, examiners were asked to complete a questionnaire comprising three parts:

- Part A sought to investigate examiner perceptions in relation to their assessment of CC
- Part B asked questions in relation to examiner perceptions of the training in CC
- Part C collected information about the background qualifications and experience of the participants.

Question types included five-point Likert scales, yes/no type questions and ranking questions (see Appendix 5). To increase the validity of the measuring instrument, the questionnaire underwent four drafts, it was piloted and discussed with four experienced IELTS examiners, including a senior examiner, and was checked by the quantitative research consultant of the university.

To investigate Research Questions 3 and 4 pertaining to examiners' marking reliability and the impact of intervening variables such as examiner qualifications and experience, Spearman correlations were calculated between the scores of each examiner on each criterion and the total scores and the standardised scores for each provided by IELTS. A confidence interval around the acceptable correlation of 0.8, as recommended by Alderson, Clapham and Wall (1995, p 132), was calculated according to the methods of Howell (1982, p 242). As Spearman correlations are not normally distributed and as they were to be used as data in further analyses, the distribution of the scores was changed using the Fisher transformation so that they were suitable to use as data in parametric hypothesis tests where applicable. To assess the reliability of examiners on each criterion, the mean correlations of the scores for each criterion were then compared using a repeated measures Analysis of Variance. The results were compared using the Bonferroni adjustment, which corrects the probabilities in the results according to the number of comparisons made. To aid interpretation, the mean scores of the Spearman correlations are reported, rather than the scores that were produced using the Fisher transformation.

To assess harshness or leniency of the individual examiners, the mean scores across all criteria for each examiner were then compared against the standard scores using independent samples t tests. The influence of a number of factors on the reliability of examiners, such as gender and years of teaching experience, were assessed by conducting independent samples t tests of mean differences in the correlations of examiners' scores and standard scores on CC. Where the number of participants in at least one group was below 15, a non-parametric independent samples test, Mann-Whitney U, was conducted. Where any means comparisons were conducted, such as ANOVA or t tests, the data was tested to assess whether the groups had equal variances, using Levene's test for homogeneity of variance. In those cases, the scores were also assessed for normality using the Shapiro-Wilk test of normality. All analyses were conducted in SPSS 13 and Systat 13.

4 FINDINGS

The data in relation to each of the research questions has been generated from both the qualitative Phase 1 and the quantitative Phase 2 of the study and will, therefore, be reported under each of the research questions.

4.1 Research Question 1

Do examiners find the marking of CC more difficult than the marking of the other three criteria?

Shaw's finding (2004) that examiners tend to find the assessment of CC more difficult than the assessment of the other criteria is supported in this study by evidence from the think-aloud process, to a lesser extent from the follow-up interviews, and more substantially from the quantitative survey results.

4.1.1 The think-aloud protocols

In the think-aloud protocols, one measure that could be taken as an initial indicator of the degree of difficulty in marking is the length of time taken to assess each criterion. Analysis of this measure, in terms of the distribution of segments devoted to each criterion, indicated that the marking of CC and TR may be more difficult than the marking of Lexical Resource (LR) or Grammatical Range and Accuracy (GRA). A higher proportion of all segments was devoted to the assessment of Task Response (TR) and CC than the proportion dedicated to the other two criteria – 24% of all segments were devoted to the interpretation and assessment of TR and 22% to CC. In contrast, 16% of all segments were devoted to LR and only 12.5% of all segments were dedicated to the assessment of GRA. If we look at the individual examiners' coded segments, more were devoted to CC than TR for six of the 12 examiners.

These findings would seem to suggest that the proportion of time spent on the assessment of TR is slightly higher than the time spent on CC overall. However, examiners spent considerable time on the interpretation of the writers' answers. If the interpretation segments are subtracted from TR, then we find that a higher number of segments were devoted to CC than TR by eight of 12 (two-thirds) of the examiners, and the overall proportion of segments for the assessment of TR is reduced to 21%, roughly equivalent to the time devoted to CC.

Another possible indicator of the degree of difficulty examiners experience in assessing the different criteria may be the amount of time they devote to reading or referring to the actual band descriptors as they assess each script. If that is the case, then more time was spent overall, in terms of the number of

segments, on reading the CC band descriptors than on reading the other band descriptors – 29% of the band descriptor reading segments were allocated to CC as compared to 28% for TR, 21 % for LR and 19% for GRA (see Table 2). However, in terms of individual examiners, only four of the 12 examiners devoted more segments to reading the CC band descriptors than to the other descriptors, while a further four examiners spent an equal number of segments reading both the TR and CC band descriptors. Least time was spent reading the band descriptors for GRA.

	Examiners with 5+ years' experience						Examiners with less than 2 years' experience						Total	
	D	M	A	F	J	P	K	T	E	S	B	L	Total	%
TR band descriptors reading segments	29	7	15	31	5	13	5	5	11	4	4	11	140	**28**
CC band descriptors reading segments	29	7	15	12	0	12	7	7	27	16	4	10	146	**29**
LR band descriptors reading segments	15	8	13	25	0	6	5	1	12	4	11	4	104	**21**
GRA band descriptors reading segments	22	5	11	9	0	10	0	3	19	3	12	2	96	**19**
Total reading segments	95	27	54	77	5	41	17	16	69	27	31	27	486	**100**

Table 2: Number of segments dedicated to reading the band descriptors

A further possible measure of the degree of difficulty in marking each criterion is the amount of hesitation, or the number of segments coded as hesitation for the assessment of each criterion in the transcripts. Initial analysis would seem to suggest that, overall, examiners were slightly more hesitant for the marking of CC than for the other criteria with 32% of all hesitation segments pertaining to CC, as opposed to 28% of hesitation segments for TR, 17% for GRA and 15% for LR (see Table 3). However, the story appears to be more complex than that. In terms of individual examiners, while Examiner P and Examiner E were much more hesitant when marking CC than for the other criteria and Examiner F was slightly more hesitant, the remaining nine markers appeared to hesitate more when marking TR.

Individual differences in marking styles were particularly noticeable for Examiner D, who was extremely hesitant and took almost twice as long to finish the marking, and Examiner B who was a very confident marker and seldom hesitated over the assessment process. Despite individual differences like these, what does seem to be clear is that significantly fewer hesitations were recorded for the assessment of LR and GRA than for the other two criteria.

	Examiners with 5+ years' experience						Examiners with less than 2 years' experience						Total	
No of Hesitancy Segments	D	M	A	F	J	P	K	T	E	S	B	L	**TOT**	**%**
TR	10	13	8	5	3	7	6	3	5	8	1	5	**74**	**28%**
CC	9	9	7	7	2	20	5	2	14	6	0	4	**85**	**32%**
LR	10	3	5	3	0	4	5	3	2	1	3	2	**41**	**15%**
GRA	16	0	1	3	0	12	3	6	3	2	0	0	**46**	**17%**
ALL	16	0	0	0	0	2	1	0	1	0	0	0	**20**	**8%**
TOTAL SEGS	61	25	21	18	5	45	20	14	25	17	4	11	**266**	**100%**

Table 3: Number of segments coded as examiner hesitancy

4.1.2 Interviews

The data from the interviews yielded some mixed findings on Research Question 1. While seven of the 12 examiners indicated that all criteria are equally difficult to assess, four examiners expressed the view that CC is the least clear of the criteria. Examiner K, for example, pointed out that CC has the longest set of descriptors, and explained that the length of the descriptors distracted the examiner's attention from the script itself. Similarly, Examiner S commented:

> *I tend to do CC last because that's the one I'm least clear about. There's a fair bit to look at there. It's easier if you look at the others first. (S)*

While Examiner J admitted that at the training course:

> *I was very confused – the thing that I had least control over was CC rather than the other criteria. I don't do very much high level teaching, so it's not something I'm looking for [usually]. (J)*

One of the examiners found CC easier to mark than the other criteria, saying that she found paragraphing made it easy for her to identify logical progression.

Data from the interviews also indicated that lack of confidence in assessing CC was an issue for some examiners. While half the interviewees were reasonably confident in marking CC ('I know overall that it will all balance out', as Examiner M said), and Examiner B had no hesitation at all, four examiners expressed uncertainty. Examiner D asserted:

> *I'm never confident, not ever. Never, never, never. I always hesitate between the, ah, criteria constantly – go backwards and forwards. I notice particularly this time that it's just a nightmare because I go backwards and forwards on the student's work, the task itself and I go backwards and forwards on the criteria and then when I'm on the next one I'm still thinking of the other one, backwards and forwards. (D)*

4.1.3 Surveys

In the Phase 2 survey, examiners ranked the four criteria from 'most difficult to mark' to 'least difficult'. Results revealed that the majority of examiners (66% n=35) ranked CC as the most difficult criterion to mark, whereas 20% (n=11) ranked TR as the most difficult. Only 4% (n=2) ranked LR as the most difficult and none of the sample ranked GRA as the most difficult to mark.

LR was ranked the easiest of the four criteria to mark by 33% of the respondents (n=18), GRA by 27% (n=15) and TR by 20% or 11 of the respondents. In contrast, only one examiner ranked CC the easiest criterion to mark. Seven examiners (13%) indicated that they considered all the criteria equally difficult or easy to mark (see Table 4).

	TR		CC		LR		GRA	
	n	%	n	%	n	%	n	%
1= Most difficult	11	20%	35	**66%**	2	4%	0	0
2	19	**35%**	11	20%	6	11%	10	18%
3	7	13%	1	2%	22	40%	23	**42%**
4= least difficult	11	20%	1	2%	18	**33%**	15	27%
5=Same level of difficulty	7	13%	7	13%	7	13%	7	13%

Table 4: Responses to the question, 'In general, which criterion do you usually find most difficult to mark?'

Most examiners indicated that they were reasonably confident with marking all four criteria (see Table 5). However, a larger number of examiners expressed less confidence in the marking of CC than in the marking of the other three criteria. While 84% (n=46) of the examiners indicated they were either confident or very confident in their marking of TR, 93% (n=51) were confident or very confident in their marking of LR and 94% (n=52) were either confident or very confident in their marking of GRA, only 60% (n=33) were confident or very confident in their marking of CC. For those examiners who were less confident, 15% indicated that they were not very confident in their marking of CC as opposed to only 4% for the marking of both TR and LR. In contrast, only one examiner indicated he or she was not very confident in the marking of GRA.

	TR		CC		LR		GRA	
	n	%	n	%	n	%	n	%
1. Not at all confident	0	0%	0	0%	0	0%	0	0%
2. Not very confident	2	4%	**8**	**15%**	2	4%	1	2%
3. Neither confident nor unconfident	7	13%	14	25%	2	4%	2	4%
4. Confident	38	69%	32	58%	39	71%	37	67%
5. Very confident	8	15%	1	2%	12	22%	15	27%

Table 5: Examiners' levels of confidence in marking each criterion

Findings from all three sources seem to support the view that a significant proportion of examiners tend to find the marking of CC more problematic than the marking of the other three criteria. We turn next to Research Question 2 to explore in more detail what examiners are looking for in their marking of CC to gain insights into why examiners tend to find CC more difficult to mark.

4.2 Research Question 2

What are examiners looking for in marking CC in Task 2? What features of Task 2 texts affect their decision-making in relation to the assessment of coherence and cohesion?

One of our original hypotheses was that examiners may not pay as much attention to propositional coherence and semantic links as they do to explicit cohesive devices in each script. However, the think-aloud data do not appear to support this hypothesis. Approximately 72% (451 segments) of the examiners' assessment of CC was devoted to coherence as opposed to 28% (176 segments) devoted to cohesion (see Table 6). This would seem to indicate that examiners as a group were spending more time focused on features of coherence in each text at the macro level than to the identification and assessment of explicit micro level cohesive devices.

Of the 72% of codes dedicated to coherence in the think-aloud data:

- 23% (147 segments) were focused on the general features or qualities of the text, such as the flow, fluency or overall clarity or coherence of the text
- 26% (162 segments) focused on aspects of logic, logical organisation, logical progression, the logical relationships or semantic links between ideas
- 23% (142 segments) were specifically focused on paragraphing, including references by eight of the examiners to introductory and concluding paragraphs in a number of the scripts (see Table 6).

Another question we wished to investigate was whether examiners were likely to pay greater attention to some of types of cohesion such as discourse markers, than to other cohesive features. The think-aloud data seems to support this hypothesis. Of the 28% of segments coded under cohesion, 20% focused on the assessment of explicit discourse markers, coordinating conjunctions or cohesive devices, terms which were used interchangeably by the Phase 1 examiners. All examiners made reference to these in the scripts under examination. All but one referred to discourse markers or sequencers a number of times with Examiners F and B referring to them a total of 11 times each. In contrast, only 5% of all codes were focused on reference and/or substitution.

Differences in the interpretation of the band descriptors for CC seemed to be evident in the emphasis individual examiners gave in their assessment of the different features of CC identified in the band descriptors. For example, there was considerable individual variation in the proportion of time spent on the assessment of coherence compared to the proportion of time spent on the assessment of cohesion. Examiner A focused 90% of the think-aloud protocol on the assessment of aspects of coherence, 7% on the assessment of 'discourse markers' or 'linking words' and noted one case of 'reference' in the marking of 10 scripts. In contrast, the segments of Examiner K focused on coherence 39% and on cohesion 61% of the time. Examiner K made 19 explicit references to the terms cohesion, cohesive devices, coordinating conjunctions, discourse markers or linking words (38% of segments). She used the terms 'reference and substitution' seven times but made only one explicit reference to 'logical organisation' and two to 'logical progression'. These findings seem to suggest that her understanding of CC was particularly influenced by more overt linguistic features of text and less by consideration of propositional coherence. The other 10 examiners ranged between Examiners A and K in the degree to which they emphasised the assessment of coherence over cohesion.

There was also some variation in emphasis between examiners in the assessment of aspects of logic. More references were made to logic, logical organisation or logical progression by the less experienced examiners than by the more experienced examiners, with Examiner M, for example, making only one explicit reference to logical organisation in the assessment of all 10 scripts.

Despite these individual differences in emphasis, however, logical organisation and paragraphing were referred to by all 12 examiners and logical progression by all except one. This would seem to indicate the importance of logical organisation and paragraphing in the decision-making of all the examiners, regardless of individual differences in marking.

Another example of a possible difference in the interpretation of the band descriptors for CC is the fact that, while eight examiners appeared to assess reference and/or substitution in a set of 10 scripts, four examiners made no explicit reference to these terms while assessing the same scripts. Differences between examiners such as these may have implications for both the reliability and the construct validity of this criterion.

The think-aloud protocols showed that all 12 examiners stayed reasonably closely to most of the features identified in the band descriptors for CC and used the terminology of these descriptors extensively while assessing the 10 scripts. However, examiners also introduced a number of other terms. These included the terms, 'flow', 'overall structure' and 'linking words', the last being used interchangeably with 'discourse markers', 'coordinators' or 'transition signals'. 'Overall structure' seemed to be used in place of 'logical organisation'. The term 'flow' was used by two thirds of the examiners. In many cases, examiners appeared to be assessing this concept in an intuitive, impressionistic way, although one examiner was clearly using the term to mean 'logical progression'. More research is needed to identify more precisely how examiners define and use the term 'flow' in their assessments.

Further features examiners noted in the Phase 1 data but which are not in the band descriptors were as follows:

- six of the 12 examiners in the think-aloud data made judgements about the 'introduction' to a script
- six examiners referred to the 'conclusion' of particular scripts in their assessment of CC
- four examiners made explicit reference to the term 'essay'
- three examiners made reference to a script writer's 'argument'
- three examiners referred to the term, 'topic sentence'
- eight examiners in the interviews referred to the term, 'topic sentence'.

Reference in the think-aloud protocols by some examiners to features not explicitly referred to in the band descriptors for CC would seem to provide further evidence for a degree of variability in the ways examiners may interpret the band descriptors for CC.

	Examiners with more than 5 years experience												Examiners with less than 2 years experience													
Features of CC explicitly discussed	**D**		**M**		**A**		**F**		**J**		**P**		**K**		**T**		**E**		**S**		**B**		**L**		**total**	
coherence	6	9%	5	22%	7	17%	3	4%	1	5%	1	3%	1	3%			3	4%			8	7%	9	23%	44	7%
meaning/message/ideas			1	4%	6	14%	4	5%			3	9%	2	6%			4	5%	1	1%	4	4%	4	10%	29	5%
argument					3	7%	4	5%													1	1%			8	1%
flow/fluency			3	13%	2	5%	4	5%			4	12%	2	6%	6	15%	9	12%	4	5%	11	10%			45	7%
clarity	1	2%			2	5%	5	7%	1	5%					1	3%	1	1%	4	5%	5	5%	1	3%	21	3%
logic	5	8%			2	5%	2	3%	1	5%					4	10%	4	5%	1	1%	3	3%			22	4%
logical organisation	9	14%	1	4%	3	7%	3	4%	3	15%	5	15%	1	3%	4	10%	6	8%	6	8%	10	9%	1	3%	52	8%
logical progression	8	13%			4	10%	8	11%	2	10%	4	12%	2	6%	4	10%	9	12%	7	9%	4	4%	9	23%	61	10%
logical rel/linking btw ideas	1	2%			1	2%	2	3%	0	0%	1	3%			2	5%	8	10%	3	4%	8	7%	1	3%	27	4%
paragraphing	20	31%	5	22%	8	19%	6	8%	5	25%	10	29%	3	10%	6	15%	15	19%	16	22%	23	21%	7	18%	124	20%
introduction	1	2%					1	1%					1	3%					3	4%	5	5%	1	3%	12	2%
conclusion							1	1%			1	3%					1	1%	1	1%	1	1%	1	3%	6	1%
TOTAL COHERENCE	**51**	**80%**	**15**	**65%**	**38**	**90%**	**43**	**57%**	**13**	**65%**	**29**	**85%**	**12**	**39%**	**27**	**68%**	**60**	**78%**	**46**	**62%**	**83**	**78%**	**34**	**87%**	**451**	**72%**
cohesion	1	2%	1	4%			11	14%	0	0%	0	0%	2	6%	1	3%	1	1%	5	7%					22	4%
cohesive devices	3	5%	1	4%			5	7%	0	0%	1	3%	3	10%	7	18%	5	6%	1	1%	7	7%	3	8%	36	6%
coordinating conjunctions	1	2%					3	4%					2	6%	2	5%	1	1%	5	7%	1	1%	1	3%	16	3%
discourse markers/link words	8	13%	6	26%	3	7%	11	14%	4	20%	4	12%	5	16%	2	5%	5	6%	9	12%	11	10%	1	3%	69	11%
reference and substitution							3	4%					6	19%	1	3%	3	4%			4	4%			17	3%
reference					1	2%			3	15%			1	3%			2	3%	8	11%					15	2%
substitution																					1	1%			1	0%
TOTAL COHESION	**13**	**20%**	**8**	**35%**	**4**	**10%**	**33**	**43%**	**7**	**35%**	**5**	**15%**	**19**	**61%**	**13**	**32%**	**17**	**22%**	**28**	**38%**	**24**	**22%**	**5**	**13%**	**176**	**28%**
TOTAL SEGS	**64**		**23**		**42**		**76**		**20**		**34**		**31**		**40**		**77**		**74**		**107**		**39**		**628**	

Table 6: Coding frequency of segments dedicated to features of CC explicitly referred to in the think-aloud protocols

4.2.1 Ranking of key features of CC: Phase 2 results

Examiners in Phase 2 ranked eight key features of CC in terms of their perceived importance (see Table 7), from 1 'most important' to 8 'least important'. In response to a second question, they indicated how often they refer to the same eight key features while marking (see Table 8). Taken together, these two survey questions are intended to provide insights into the salience of these features in the examiners' perceptions of CC.

The terms 'reference', 'substitution', 'paragraphing', 'message/ideas' and 'logical progression' were included in the list of features of CC examiners were asked to consider because they are key terms in the band descriptors for this criterion. The terms 'linking words', 'flow/fluency' and 'overall structure' were included as features of CC because these terms were frequently referred to by the examiners who participated in the qualitative first phase of this study, even though these particular terms are not used in the existing band descriptors.

RANKINGS	1		2		3		4		5		6		7		8	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Reference	1	2	1	2	5	9	7	13	12	22	13	24	8	15	7	13
Substitution	0	0	1	2	4	7	6	11	11	20	7	13	18	33	7	13
Paragraphing	**6**	**11**	3	5	10	18	6	11	8	15	11	20	6	11	4	7
Message/ideas	5	9	7	13	7	13	5	9	5	9	4	7	8	15	**13**	**24**
Linking words	4	7	12	22	8	15	12	22	9	16	6	11	3	5	0	0
Flow/fluency	**15**	**27**	12	22	6	11	6	11	4	7	3	5	5	9	2	4
Overall structure	**6**	**11**	5	9	6	11	7	13	2	4	9	16	4	7	**15**	**27**
Log Progression	**17**	**31**	13	24	13	24	4	7	4	7	1	2	1	2	1	2

Table 7: Examiners' rankings of features of CC in terms of their importance in the assessment process

Examiners ranked the following features of CC in <u>either first or second</u> position:

- logical progression (55% n=30)
- flow/fluency (49% n=27)
- linking words (29% n=16)
- message/ideas (22% n=12)
- overall structure (20% n=11)
- paragraphing (16% n= 9)
- reference (4% n= 2)
- substitution (2% n= 1)

Features ranked in the last two positions in terms of their relative importance, were as follows:

- substitution (46% n=25)
- message/ideas (39% n=21)
- overall structure (34% n=19)
- reference (28% n=15)
- paragraphing (18% n=10)
- flow/fluency (13% n= 7)
- linking words (5% n= 3)
- logical progression (4% n= 2).

Although the ranking exercise forced examiners to identify priorities somewhat artificially, these findings suggest that some examiners may have slightly different perceptions of the importance of certain features and the role they play in the marking of CC. Most agreement can be found in the rankings of logical progression and substitution. Logical progression seems to play the most important role with 55% (n=30) of examiners ranking this feature in either first or second place, with only two examiners ranking this feature in the last two positions.

At the other end of the scale, 'substitution' was ranked in last or second last position by 46% (n=25) of examiners. This finding is in line with the work of Halliday and Hasan (1976) who noted that the use of substitution is rare. However, it may also be that 'substitution' has been ranked last in a few cases, because the concept appears to be poorly understood by a number of examiners, as indicated by the definitions they supplied (see section 4.2.6).

However, 'flow/fluency' was also ranked in the first two positions by almost half the examiners (49%), while only four examiners ranked this feature in the final two places, despite the fact that the term 'flow' is not used in the band descriptors and does not lend itself very readily to analytical assessment. It may be that 'flow' is another term that some examiners use interchangeably with 'logical progression' as one examiner noted. As noted in the previous section, analysis of the think-aloud data suggests that some examiners tended to assess 'flow' intuitively, with little indication that they were analysing the logic or logical progression of ideas.

There appeared to be less agreement about the relative importance of other features of CC in the assessment process. For example, paragraphing was ranked in the top two places by nine of 55 examiners but it was also ranked in the last two places by 10 examiners. A possible explanation for the range of responses in relation to paragraphing may be the differing perceptions of a number of examiners over the role played by the paragraph ceilings (see section 4.2.3). Examiners also gave a range of responses in the ranking of 'message and ideas'. While 12 examiners ranked this feature in the top two positions, 21 examiners placed message and ideas in the bottom two places. Similarly, although 'overall structure', not a term used in the descriptors, was placed in the bottom two positions by 19 of 55 examiners, 11 examiners placed 'overall structure' in the top two positions.

Participant examiners also identified the frequency with which they refer to the same eight key features of CC in their assessments. The overall results (see Table 8) support the general findings of the previous ranking exercise in placing 'logical progression' and 'flow/fluency' as those features most often referred to and 'substitution' and 'reference' least frequently referred to.

Differences of opinion over the role played by 'message/ideas' as well as 'overall structure' seem to be borne out by the wide distribution of responses for these two features – 26% (n=14) of examiners always refer to 'overall structure', 36% (n=19) refer to it very often, 26 % (n=14) refer to it only sometimes and 11% (n=6) seldom. 'Paragraphing' on the other hand, was referred to very often by 23% (n=12) and always by 52% (n=27) of the examiners.

	never		seldom		sometimes		very often		always	
	n	%	n	%	n	%	n	%	n	%
Reference	0	0%	6	11%	20	**37%**	19	36%	8	15%
Substitution	0	0%	7	13%	25	**47%**	18	34%	3	6%
Paragraphing	0	0%	2	4%	11	21%	12	23%	27	**52%**
Message/ideas	3	6%	5	10%	13	25%	18	**35%**	13	25%
Linking words	0	0%	0	0%	9	17%	23	**43%**	21	40%
Flow/fluency	0	0%	2	4%	8	15%	13	25%	30	**57%**
Overall structure	0	0%	6	11%	14	26%	19	36%	14	26%
Log Progression	0	0%	0	0%	6	11%	15	28%	32	**60%**

Table 8: Examiners' perceived frequency of use of features of CC

In the next section, we give more detailed feedback on examiners' perceptions of the key features of CC.

4.2.2 Coherence

Think-aloud protocols

Examiners appeared to divide their time fairly equally between the three key areas of coherence: 23% of codes were dedicated to the comments about coherence or the general qualities of each text such as flow, fluency and clarity as examiners gained an overall impression of the text; 26% of codes were devoted to aspects of logic; and 23% to paragraphing.

Typical comments in relation to the general qualities of the text include the following:

- It's just a bit incoherent. (A, line 88)
- There's coherence there. (A, line 362)
- Not sure that makes sense. (T, line 108)
- There is a good level of clarity and fluency in this piece of writing on the whole though. (E, line 90)
- It's not fluent. There lacks fluency. (F, lines 34-35)
- Well the sentences sort of flow on nicely. (P, line 243)
- But the fluency and logic flow is not clear. (B, line 47)
- So it's quite easy to go through and follow what the person is saying. (B, line 74)
- It's pretty good, it flows quite well. (S, lines 99-100)

In general, examiners in the think-aloud data did not provide concrete evidence with which to support their impression or intuition about the coherence of the text, although in some cases they referred to paragraphing for this purpose.

Three of the more experienced examiners, A, M and F, and two of the less experienced examiners, K and L, devoted more than 40% of their 'coherence' assessment time in gaining an overall impression of the scripts in terms of their general qualities such as the flow, fluency, clarity or general coherence, perhaps suggesting that they tended to be more impressionistic in their marking.

Some examiners tended to emphasise features that are traditionally associated with the structure of argumentative essays (introduction, body, conclusion, topic sentence and so on); terms which are not, in fact, used in the CC band descriptors. Take, for example, this extract from Examiner B's protocol:

209 Well the good thing is that there is presentation of very clear paragraphing
210 albeit the introduction is a single sentence.
211 The 2 body paragraphs begin with 'first' and 'second'
212 and the conclusion begins with 'eventually', spelt correctly.
212 So paragraphing is there
213 and there is an element of logicality to the paragraphs.

(B, lines 209-213)

Interviews

When asked what they thought coherence meant, eight of the 12 examiners in the Phase 1 interviews, characterised coherence principally as 'flow', 'fluency' and 'clarity'.

> *Coherence is if it sort of flows okay; then I would say it was coherent. (T)*

> *Coherence I always think is about my understanding of what you mean. So it's your clarity. It's your, your sort of strategic choice of vocabulary that's going to get the message across. So it's about fluency and clarity in your style of writing. (E)*

> *Look at how it FLOWS really nicely! (P)*

Other examiners tended to see coherence as being characterised by rhetorical structure and argument. While paragraphing was important to these examiners as it was to all participants, they looked particularly for logical organisation and argument. In the interviews, these examiners talked about coherence in the following terms:

> *Coherence? I generally read through the thing and there's a developed argument and it's paragraphed. Cos the visual thing is important for me. If there's paragraphing and I also get a sense of the argument that go together, and it looks well-organised. So it's the organisation and the development of the argument – that's coherence. (A)*

> *Coherence to me is the overall idea, it's the organisation of the ideas, like for example, is there an introduction, is there a conclusion? And are there points and do they follow? (F)*

> *Coherence – one's looking at paragraphing. And I'm expecting the introduction to tell me what we're talking about, what the writer's going to attempt to achieve. Then looking for clearly delineated body paragraphs which are going to address what – hopefully – the introduction advertised as being the main points that are going to be talked about … (B)*

Examiners who seemed to focus more on structure tended to define coherence more in terms of 'logic' and 'logical organisation'. Examiner B, for example, a very systematic marker, talked of 'a logical stepping arrangement'. Examiner S provided a more detailed explanation of 'logic' suggesting examples of logical organisation as moving from general to specific, or using chronological ordering of ideas. One examiner, J, pointed out that logic is a cultural construct, and that she needed to look carefully for the logic of candidates' answers.

> *Logical order – making sense. But some people think in different ways so their logic is different from my logic so you sometimes have to think again that you can't just take it at face value that you put this first that you can't make a logical argument without that. (J)*

Overall, the Phase 1 think-aloud and interview data seem to indicate that when assessing coherence, the examiners tended either to fall back on a holistic, intuitive impression of the text in order to assign a grade, or alternatively to invoke structuralist understandings of coherence derived from their understanding of the traditional essay genre.

Surveys

Results from the Phase 2 surveys indicated that there was a relatively high level of agreement about the meaning of 'coherence'. In general terms, most examiners agreed with Shaw and Falvey's (2008, p 42) definition of coherence as 'conceptual'. Forty-five examiners (80%) define coherence in terms of the 'clarity', 'comprehensibility', 'understandability' or 'intelligibility' of the 'meaning', 'message' or 'ideas' conveyed by a text. Approximately half of these examiners referred to making sense of the 'meaning', 16 to the understandability or development of 'ideas', and seven defined coherence in terms of the clarity of the 'message'. Twenty-five examiners (45%) indicated that some kind of 'logic' was necessary to ensure clarity of meaning. Seven examiners referred specifically to 'logical progression', 'logical sequencing' or 'logical development', indicating perhaps a perception of writing as a dynamic, developmental process, while 33% (n=18) used the terms, 'logical organisation' or 'logical structure', possibly conveying a more static or structural approach to coherence in text.

Explicit reference was made to the deductive essay format by 9% (n=5) of examiners, who referred especially to the necessity for an argument with an introduction, body and conclusion structure, and with paragraphs containing topic sentences together with supporting ideas and evidence. Four examiners (7%) used the term 'essay', despite the absence of this term from the IELTS band descriptors.

Eight examiners (15%) referred to the term, 'flow' as in, 'the flow of the text' or 'flow of ideas'. That 'flow' is a term which may sometimes be equated with 'progression' is suggested by one examiner who supplied the following definition, '[Coherence is the] logical flow or progression of an argument or sequence of ideas which put together, support a stance/an argument'. However, it is not clear whether all examiners who used the term would necessarily equate 'flow' with 'progression'.

A number of examiners indicated the difference between coherence and cohesion as being between the macro and micro organization of text suggesting, perhaps that coherence refers to the overall picture and cohesion to the details needed to achieve coherence. Other definitions which would seem to support this idea of coherence as the bigger picture included the following:

- overall organisation/logic of text
- the overall comprehensibility of a text
- overall logical development of a text
- coherence refers more to overall structure of a text as a whole
- the overall structure of a piece of writing
- the extent to which the writing as a whole makes sense
- …the overall meaning of the writer's answer is clear
- the overall combining of the parts to provide a meaningful whole

- overall clarity of the script
- to do with the overall message
- overall understandability
- refers mainly to whole text structure
- overall clarity of expression.

Although there was a relatively high level of agreement for the meaning of coherence in broad terms, in some cases examiners were less clear about the more precise meaning of coherence and its relationship with cohesion. For example, in the following definitions, coherence appears to have been conflated with cohesion:

- 1) structure of the paragraph = topic sentence + supporting and developing sentences
  2) logical progression of ideas using linking words

- sentence structures are clear and linking words are used. Appropriate substitution is used and lack of repetition. Each paragraph has a clear idea.

4.2.3 Paragraphing

Think-aloud protocols

Paragraphing contributed 20% of total codes and was the most used code in the think-aloud data. It was also one of the first features of text that caught examiners' attention. For example, Examiner M as she began to look at Script 9 commented as follows:

148 Right, moving on to number 9.
149 OK, which is incredibly short.
150 OK, this will be minus 2 on the TR.
151 I'm looking already
152 there's no paragraphs
153 there's just one large paragraph
154 so that's going to be a problem with coherence. OK.
(M, lines 148-154)

Three examiners discussed good paragraphs in terms of topic sentences plus supporting sentences – see the extract from Examiner S's transcript below.

106 The only problem I find is that his paragraphs are relatively short.
107 They might be only two sentences long
108 so effectively you've got a topic sentence and then a supporting sentence
109 but the supporting sentence doesn't really fit with the topic sentence.
110 It kind of it sort of jumps from the topic sentence to the supporting sentence
111 and it doesn't really blend.
(S, lines 106-111)

Although the band descriptors mention the concept 'clear topic', neither they nor the training materials mention the term 'topic sentence'. Examiners who used these terms also tended to refer to the script as an 'essay' rather than using the IELTS terminology of 'response' (For more on this issue, refer to section 4.3.2).

The same underlying understanding of effective paragraphing as a feature of logical organisation appears to underlie Examiner A's thinking in the following excerpt.

206 Well first of all there are paragraphs
207 this thing has four paragraphs – an introduction, two bodies and a conclusion.
208 It's organised.
209 This is tricky.
210 I think this is a 5.
211 4 says (reads aloud)
212 There are clear main topics and there are paragraphs.
213 Yeah, I think this is a 5 on that respect.
(A, lines 206-213)

Interviews

Data from the interviews provides further evidence of examiners' use of the term 'topic sentence'. While the descriptor term 'clear central topic' was mentioned by only three examiners, the term 'topic sentence' was used by eight of the 12 examiners in Phase 1 interviews.

> *Well, there should be a topic sentence and that should control the paragraph and, um, so, yeah, I don't think that's really, paragraphing is skilfully managed. Well, if its skilfully managed there would be a topic sentence and the supporting sentences, which support the topic sentence. (M)*

> *...paragraphing ... is there a structure like a, well, the conventional one of a topic sentence, elaboration and examples. (L)*

The interview data also showed that paragraphing was an aid to navigating the scripts.

> *Well, I think it's pretty clear – having one central topic for each paragraph and referring back – linking back to the paragraph before and so on. So paragraphing I find – like 'no paragraphing and/or clear topic within paragraphs'. That I find quite easy to see. (P)*

> *I actually think that I go from paragraphs – having read the thing and decided either it's logical or it's not – Are there paragraphs? How are the paragraphs structured? What's the start of each paragraph? So I'll go from the paragraphing. And then after I've looked at the paragraphs I'll look to see if they are structurally meaningful paragraphs. (B)*

However, some examiners felt that they were not clear about what was expected in terms of paragraphing.

> *And I don't think I've ever really got a clear answer to what constitutes a paragraph that's um consistent in my training. (M)*

Several examiners pointed out that superficially, cohesive paragraphs were not always very meaningful.

> *And the paragraphing [...] it can be beautifully done, you know each paragraph is introduction – firstly – secondly and then finally – in conclusion. It's all very well. Superficially it looks terrific but it's not always worth very much. (S)*

A number of examiners expressed reservations about the paragraph ceilings.

> *So paragraphing is important? Yes [...] but sometimes I get upset when you have got someone who has written beautifully but hasn't put in appropriate paragraphs whereas the message is there, or where there is a little missing line or indentation. But I follow my instructions. (J)*

> *I remember a native speaker I had once with no paragraphing at all. That was kind of extraordinary. And I knew it was a native speaker because of the language and everything else and yet I couldn't, it was really awkward, I just couldn't give her or him a 9 because of*

the paragraphing so… Um, you know, if paragraph is missing then where is it? It's way down at 5 isn't it? "Paragraphing may be inappropriate or missing". (D)

There's no paragraphing and I don't think that there's any "firstly", "secondly", there's no coordinators that you know give style to any of the sentences but it's all right. I can't see anything wrong with it. Not really anyway. It's fairly logical. It's one train of thought. It just goes on and on and on. It's not paragraph one-idea one; paragraph two-idea two; paragraph three-idea three linked cleverly. But it doesn't mean that it's any worse. (D)

These examiner observations are similar to the findings of Shaw (2006, p 5). Examiners surveyed in the trial study of the new descriptors noted that paragraphing may have been given too much emphasis in the revised scales so that candidates who have not been taught paragraphing might be disadvantaged.

Surveys

Examiners were asked to indicate the usefulness of the bolded ceilings for paragraphing in their assessment of Academic Task 2 texts (Survey Question 13) – 13% (n=7) rated the ceilings as not very useful, 45% (n=25) as quite useful, 24% (n=13) rated them as very useful and 18% (n=10) as very useful indeed. As indicated previously, some examiners in the interviews expressed concern about the requirement to award a score no higher than 5 for a piece of writing with no paragraphs, even though there are instances of testees writing relatively coherent and cohesive texts but without each new thought set out in a separate paragraph. It may be that the relatively low priority accorded by some examiners for paragraphing in the ranking exercise, as shown in Table 6, reflect this view.

4.2.4 Cohesion

The interview data and the definitions examiners provided in the surveys revealed that, on the whole, there was a reasonable level of agreement about the general meaning of cohesion. In the surveys, just over half of the examiners (58%, n=32) defined cohesion in terms of the methods used to link (connect/put together/arrange/relate/glue/stitch/stick/join/bond/hang together) parts of a text. Five respondents (9%) defined cohesion in terms of the ways in which a text 'flows', while eight examiners (15%) defined cohesion in terms of grammatical or syntactic methods of increasing both 'linking' and 'flow'. On the other hand, eight respondents (15%) defined cohesion with particular reference to the structuring of the text, or as one examiner put it, to 'the elements that contribute to the structuring and organisation of the text such as discourse markers and pronoun substitution'. One examiner appeared to confuse the term 'cohesion' with 'coherence' in his/her definition, and the final examiner defined cohesion simply as 'the use of cohesive devices' without further reference to their purpose.

Twenty-two of the 55 Phase 2 examiners (40%) included in their definitions that cohesive links or connections happen at the word, sentence or paragraph level. Seven (13%) referred to the ways in which ideas are connected within sentences, 12 examiners (22%) referred to the connections between sentences, and four examiners each referred to the linking within or between paragraphs.

With reference to that which is being connected within a text, 25 Phase 2 examiners referred specifically to the linking of 'ideas', one to the way in which 'information' is linked together logically, one defined cohesion as the linking of 'meanings' and two others refer to cohesion as the way 'the argument' is bound together.

Examiners were less specific about the ways in which such links are made. Fourteen examiners (25%) listed some kind of connectors (linking words, discourse markers, conjunctions, transition signals) as the main means of linking parts of a text together and 15 (27%) indicated that 'reference' was also important for this purpose. Nine examiners also listed 'substitution' together with 'reference'; two examiners mentioned the role of 'ellipsis' as a cohesive mechanism. One examiner each made specific reference to 'articles', 'synonyms' and 'topic sentences' as ways of creating cohesion in text.

In the interviews, examiners generally looked upon cohesion as a feature at the 'micro' level, as Examiner B put it, as opposed to coherence being a feature of the text as a whole. As with the assessment of coherence, individual differences in approach were apparent between those examiners who appeared to mark more analytically and systematically and those who were more intuitive in their approach. For example, the more intuitive markers tended to describe cohesion in rather general terms as in the follow extracts:

> *... the sticking together of paragraphs and sentences and ideas and notions so that it's not higgledy piggledy absolute gibberish. (E)*

> *Cohesion is I guess to some extent, the flow of it. How well it all goes together but I hadn't really – I think I focus – for myself I focus more on the coherence and if the cohesion comes into it, it's not recognising it on its own. (L)*

The most systematic examiners, F and B, referred to almost all the features of cohesion in the think-aloud marking process, both general and specific. In the interviews, they stressed the importance of the structure of paragraphs as a feature of cohesion.

> *Cohesion for me is more at the paragraph and sentence level. Looking at the structure of the paragraph, is there a topic sentence, is it introduced with a cohesive device like 'firstly', 'secondly', or 'on the other hand' or whatever. (F)*

> *When it comes to cohesion, looking within the paragraphs. We've got a topic sentence, so we know what this paragraph is going to address, and we can see that the supporting ideas are coming out, with evidence and that they are related to each other so the next sentence is a supporting idea. (B)*

4.2.5 Cohesive devices/sequencers/discourse markers

According to Halliday and Hasan (1976), the term 'cohesive devices' can generally be understood to refer to all those linguistic mechanisms providing cohesion in text, specifically: reference, substitution, ellipsis, conjunction and lexical cohesion. The examiner training materials include a slightly different range of textual features: sequencers, linking devices, and referencing and substitution under this heading. While all the examiners talked about 'cohesive devices' in both the think-alouds and the interviews, in nearly every case, the term was used synonymously with discourse markers with no mention of the other types of cohesive device.

Think-aloud protocols

Data from the think-aloud protocols showed that examiners accorded a great deal of prominence to 'cohesive devices', 'discourse markers' or 'coordinating conjunctions', which together accounted for a further 20% of all codes. For two less experienced examiners, K and T in particular: these were the most used features accounting for 32% of all K's codes and 28% of T's codes (see Table 6). Sixteen per cent of K's coding incidences concerned discourse markers, 6% referred to coordinating conjunctions and 10% related to cohesive devices. Eighteen per cent of T's segments referred to 'cohesive devices' and 5% each to coordinating conjunctions and discourse markers.

On the whole, the more experienced examiners paid slightly less attention to these explicit markers of cohesion than the less experienced examiners (see Table 6). The experienced examiners as a group made 47 references to 'cohesive devices', 'discourse markers' or 'coordinating conjunctions'. Examiner A, one of the longest-serving examiners, made only three references to discourse markers as the only aspect of cohesion assessed and the rest of the think-aloud transcript was focused entirely on the assessment of coherence. In contrast, the less experienced examiners referred to cohesive devices, discourse markers or coordinating conjunctions 71 times between them. All six inexperienced examiners referred to all three features, whereas only two of the six experienced examiners referred to all three terms separately.

Reasons for this difference between the more and less experienced examiners are a matter of conjecture. It may be that more experienced examiners are cognisant of the fact that coherent texts do not have to depend on explicit grammatical or lexical markers for their meaning to be apparent, while less experienced examiners tend to look for these explicit markers because they are more easily identified. Alternatively, the difference between the two groups may lie in the fact that the more experienced examiners – all but one of whom had 10 or more years' experience as examiners – can still recall the holistic marking scale of former days and are influenced accordingly, whereas the newly trained examiners have been taught to be more analytical, systematically considering each feature of CC in turn. More research is needed to explore these issues further.

Interviews

In the interviews, the terms 'sequencing words', 'linking words', 'transition signals' and 'discourse markers' were used interchangeably. Examiners commented:

> *I gave it a 5 for CC because there's very few. You know, the linking words are very simplistic just 'because' and 'but' and there's no, you know, any sequencing, you know, any paragraph sequencing (M)*
>
> *Sequencing – I guess the "first", "second" sort of use, that type of thing (L)*
>
> *Cohesive devices, but that's the transition signals and stuff isn't it? (P)*
>
> *Cohesive devices 'although', 'but', 'so', that's how I understand (L)*
>
> *The cohesive devices? The little words and the little tie ins (S)*

One interviewee took a slightly broader definition of 'cohesive devices'.

> *Cohesive devices would be use of link words for example, or um a good topic sentence (D)*

Several examiners pointed out that the use of cohesive devices did not necessarily indicate coherent writing.

> *Cohesive devices – well – I expect that's looking for some of those terms that we often find in our marking that students can make a beautifully cohesively devised essay with no content to it at all. They have learnt 'firstly' and 'so' and 'in addition', so they used all these words but if there is no substance to them they don't mean very much. (J)*
>
> *Often candidates have a bit of a handle on cohesive devices. Even basic ones, you know 'and', 'so', 'therefore'. They seem to get a handle on that […] Occasionally – not often – they throw in phrases that they know are supposed to act to pull it all together and it might be in the wrong context but they know they should be using these phrases. But that doesn't happen that often. The simple ones, they often seem to have a handle on that ahead of other aspects of writing. (L)*

Cohesive devices were referred to by all examiners in the Phase 1 think-aloud protocols irrespective of their marking style. In the interviews, the more systematic analytic examiners indicated that they look for discourse markers as an indication of a well-structured text.

> *Then getting down to a conclusion which tells me that we're in the conclusion because it begins with 'In conclusion' 'In summary' or with another sort of indicator which is going to define the answer to the question asked. (B)*

Survey definitions

Respondents for the Phase 2 survey were asked to provide a short definition or list of cohesive devices. Responses varied but in most cases fell short of the list provided in the examiner training materials. In general, cohesive devices were thought to be synonymous with discourse markers.

While four examiners gave vague definitions of the term 'cohesive devices' which were difficult to categorise and provided no examples to make their definitions clearer, 65% (n=36) of examiners defined cohesive devices as some form of linking word or phrase. They variously referred to these as: 'linking words or phrases', 'sequencers' or 'sequence markers', 'logical connectors', 'transition words' or 'signals', 'discourse markers' and 'conjunctions'. Twenty-two of these 36 examiners gave examples of these words or signals.

Six examiners (11%) defined cohesive devices in terms of both logical connectors or linking words and one other feature of cohesion: four of these included 'reference' in their definition, one included 'lexical chains' and another included 'paragraph structure' as a cohesive device. One examiner indicated that 'cohesive devices' was the overall term for reference and substitution. A further four examiners defined cohesive devices in terms of logical connectors or linking words together with both reference and substitution, and another two made specific mention of 'synonyms'. Two examiners gave longer definitions as follows:

- logical connectors, reference, substitution, lexical chains, theme/rheme
- linking words and transition markers, conjunctions, ellipsis, paragraphing, referencing and substitution.

4.2.6 Reference and substitution

Think-aloud protocols

The least used aspect of the CC descriptors in the think-aloud data was 'referencing and substitution'. They accounted for 5% of all codes in the think-aloud transcripts, which is a similar finding to Halliday and Hasan's (1976) analysis of a selection of sample texts. As stated earlier, four examiners did not refer to either concept in their assessment processes. Of the remaining eight examiners, five of them conflated the two terms, always using them together. 'Reference' was discussed on its own by five examiners and 'substitution' as an independent feature of CC was referred to once in the think-aloud transcripts (see Table 6).

Some confusion over definitions of referencing and substitution is apparent in the two extracts below from the think-aloud data of the same examiner. In the first, she refers to substitution as synonymy, while in the second she conflates substitution with referencing.

> *She's, well, she's summarising what she said in the first paragraph but not substituting it with synonyms or parallel expressions. (E, line169)*

> *Good examples of substitution. 'It is certainly true that products are bought anywhere on the globe and even manufactured as well but THIS does not make …' (E, line 60)*

The anaphoric use of the definite article was identified as a sub-feature of reference once in the 12 think-aloud protocols. In all other instances of the coding of reference and substitution, examiners were identifying anaphoric pronoun use, the use of synonyms or the repetition of key nouns.

Interviews

Generally in the interviews 'reference' was thought to mean the use of pronouns, while 'substitution' referred to paraphrasing.

> *The reference um well I think of as 'it' 'this means' – that sort of thing. Pronouns. (J)*

> *Substitution means … paraphrasing […] It means saying the same thing in different words. (A)*

However, several examiners in the interviews appeared to be confused about the definitions of reference and substitution and how they should be applied in marking CC. As one examiner commented:

> *'Use of reference and substitution'. That doesn't mean very much to me. 'Can result in some repetition or error' Umm. No. that doesn't mean anything to me. (P)*

Another examiner conflated the two terms.

> *Reference um, reference and substitution I'm inclined to put together in the sense of using pronouns to refer to something that's mentioned earlier. (L)*

Two examiners noted that they tend to look at reference and substitution last in their assessment of scripts.

> *The reference I don't look at too much unless it's glaringly standing out. I don't sort of find that as something I think about too much. (F)*

> *I would probably put referencing as the bottom of the things that I look for and substitution. Well if there's NICE substitution or paraphrasing, then I tend to notice and think – you know – that's good. (A)*

Surveys

Most examiners defined reference in the very broadest terms as lexical items used to refer to other lexical items in order to avoid repetition. Thirteen examiners (24%) defined the term only in these very broad terms. However, 39 examiners (70%) made specific reference to the use of pronouns to replace lexical items either earlier or later in the text. Twenty-one (38%) respondents defined reference in terms of referring back in the text: two examiners in particular used the term 'anaphoric' reference in relation to this point. Nine examiners also defined reference in terms of referring forward. One of these also used the term 'cataphoric' to indicate that lexical items might in some cases be used to refer to something further on in the text and one other examiner gave an (incorrect) example of exophoric reference.

Of the 39 examiners who defined reference in terms of the use of pronouns, 27 (49%) provided examples of personal pronoun use, 13 (24%) gave examples of demonstrative pronouns and eight (15%) provided examples of relative pronoun use. Only three examiners mentioned the use of articles as examples of reference, despite the inclusion of the definite article as one feature of reference in the examiner training materials. One examiner gave the example, 'Similar(ly)'as a comparative term which is included by Halliday and Hasan (1976, p 333) as a feature of reference. Of the remaining examiners, one simply stated, 'give examples', but it wasn't clear whether she was providing an unusual definition or an exhortation to IELTS to provide examples of 'reference' and 'substitution' in the training. Another indicated she equated reference as being the same as substitution, while another wrote, 'Pronouns and stuff like that? It's unclear'. One examiner did not provide a definition.

Definitions of the term 'substitution' seemed to indicate that some examiners were not entirely clear about the meaning of this term, and were not sure of the difference between the terms 'reference' and 'substitution'. As one examiner put it, 'I have to be honest and say these [reference and substitution] are indiscernable (sic) to me'. Definitions of 20% (n=11) of examiners indicated that substitution is a mechanism to avoid repetition but were vague about what form this mechanism takes. Definitions of this kind were along the lines of the following, 'not repeating words over and over but substituting terms and ideas coherently'. Almost half of the examiners (47% n=26) defined substitution

specifically as the use of synonyms to replace nouns in order to avoid repetition, with 11 of those examiners explicitly using the word 'synonym', and the remainder defining substitution in terms of replacing nouns with other nouns as a means of avoiding repetition. A further four examiners defined substitution as 'paraphrasing'. Another 47% of examiners defined substitution as simply the use of pronouns to replace nouns, with 13 of these examiners explicitly referring to pronominals and eight of them giving examples of demonstrative pronouns. Two examiners gave examples of substitution in line with the definitions provided by Halliday and Hasan. Both gave examples of verbal substitutions as follows:

> John went home and Mary **did** too.
>
> Are you going? I think **so**.

Of the remaining responses, one examiner each identified articles, determiners and quantifiers as examples of substitution.

4.3 Further issues in assessing the features of CC

4.3.1 Overlaps in the assessment of the band descriptors

There were a number of overlapping segments between the different criteria identified during the think-aloud analysis, where examiners appeared to hesitate over which features should be assessed under which criterion. While a certain amount of overlap is inevitable, there were a small number of significant overlaps between TR and CC, and to a lesser extent between LR and CC.

Overlaps between TR and CC

All 12 examiners recorded at least some overlap between TR and CC, which accounted for 134 segments or 3% of the total.

As Examiner A noted:

> *So I do look for coherence in the argument – and that also comes into task response. I've already sort of assessed that they've responded to that task because I understood the development of their argument. So I've already looked at that. It overlaps doesn't it? Because TR is good, that indicates that it's a coherent answer. (A)*

Two sources of ambiguity related to the wording of the band descriptors appear to account, at least in part, for the larger number of overlap segments between TR and CC than between the other criteria. The first of these would appear to be examiner difficulty differentiating between:

- how clearly the position is presented and supported in each script (to be assessed under TR)
- how clearly the message or ideas are conveyed and logically organised (assessed under CC).

Conceptually, it is difficult to see how these two key descriptors can be regarded as distinct, since one would seem to imply the other, at least to some degree.

The survey data provides some evidence for such ambiguity. Although 35% (n=19) of examiners ranked 'message/ideas' in the first three places in terms of their importance in the assessment of CC, 46% (n=21) placed this particular feature in the bottom two positions in terms of importance in their marking decision-making. Although the content words, 'message' and 'ideas' feature under CC in the band descriptors (see bands 1, 2, 8 and 9 for use of the term 'message' and bands 3 and 7 for 'ideas'), some examiners seem to consider these features to be part of the marking of TR. As one examiner explicitly noted, she never refers to 'message/ideas' as part of CC because she considers it 'part of TR'.

The second related source of ambiguity was examiner difficulty in differentiating between:

- 'the development of ideas' or the lack thereof in TR
- 'the logical progression of ideas' in CC.

The following extract from Examiner T directs attention to this issue:

114 Um, yeah, organisation. Coherent.
115 **A lack of overall progression, and that's kind of not really developing it.**
116 **But if we've already talked about in TR, clearly I shouldn't penalise her for that.**
117 Her, this one I've decided it's a girl,
118 I wonder why I think that.
119 But they're not firmly linked?
120 They ARE firmly linked.
121 Ok, more than a 5.
122 *Clear overall progression*.
123 *Cohesive devices are used to some good effect*.
124 Definitely – better than some good effect.
125 Referencing and substitution is 5.
126 And they'll always be logical. Yeah.
127 So clear overall progression, or a lack of progression… increases…
128 I'm going to go for a 6
129 **because I think I've already marked her down for not really developing it in TR**
130 **and the reason I want to go down to a 5 for CC is exactly the same thing, that there's not really much development.**
131 But I'm not really sure.
132 But yeah, I don't think I can drag her down twice.
(T, lines 114-132)

This ambiguity appears to account, at least to some extent, for the overlapping TR/CC segments where it was difficult for the researchers to decide whether the examiners were assessing TR or CC separately or together.

The definitions some Phase 2 examiners gave for 'coherence' seem to point to another possible cause for these overlaps. Five definitions of coherence included the 'relevance' of the answer to the question and whether the response was 'on task':

- Relevance – more to do with the logical presentation of ideas, addressing a question with suitable/appropriate – the understandability of an essay
- making sense. Being logical and relevant
- … does it make sense/is it relevant
- readable, clear message/communication, logical organisation, on task
- linking of essay to test topic…

It can be argued that 'relevance' is a feature better assessed under TR rather than CC, and its inclusion in examiner definitions of coherence may be evidence of the difficulty some examiners have in differentiating between some aspects of CC and TR.

Overlaps between CC and LR

Overlaps between the other criteria were not significant in terms of number. Seven examiners recorded an average of three CC/LR overlapping segments each. These seemed to relate to lexical chains, the use of synonyms and the repetition of key nouns – what Halliday and Hasan (1976) refer to as lexical cohesion – and whether such reiteration is a cohesive device to be assessed under CC or as an indication of lack of flexibility in vocabulary use under LR.

In one script, the candidate starts each of the three paragraphs in the body of his/her text with the same phrase 'the accessibility to products …'. One examiner interpreted this pattern as a fairly sophisticated theme/ rheme progression along the lines of a-b, a-c, a-d. However, a number of examiners, while acknowledging that the writing 'flowed', talked about this repetitive pattern in slightly disparaging terms under LR. For example, Examiner T seemed to doubt the effectiveness of this repetition as a cohesive measure with the use of 'although' and 'again' in line 91.

83 The vocab's better than the grammar I think
84 and that gives it the real nice flow rather.
85 (reads aloud from para 6)
89 The good noun phrase, 'the accessibility of many on the globe to products' which I guess is pretty complex.
90 'The character of countries', 'accessibility to the same product'…
91 **although then, yeah, 'accessibility to products' again.**
(T, lines 83-91)

That Examiner J in the extract below perceives reiteration to be a negative feature while at the same time implicitly acknowledging its use as a cohesive device is indicated by the use of 'but' in line 62.

56 Funny, 'that should be celebrated' he's got that right
57 but then later on 'which we need celebrate'
58 he's been repeating there – same word –
59 but not used correctly.
60 I think his spelling is generally very good.
61 **He has repeated words quite a lot**
62 **but it's quite easy to read.**
63 I think it's -
64 but there's not much in the way of uncommon words
65 so make that a 7 [for LR].
(J, lines 56-65)

The overlap between LR and CC in relation to lexical cohesion is also evident in the following extract from Examiner E while assessing CC:

168 [It's] repetitive,
169 she's, well, she's summarising what she said in the first paragraph
170 but not substituting it with synonyms or parallel expressions.
(E, lines 168-179)

In the interviews, Examiner K explained this overlap as follows:

> *The one I'm least confident about is LR because it overlaps so broadly with the other areas. So it's really hard to have a good TR with something if you don't know the vocabulary to discuss it. It's really hard to do the things that they ask you to do for CC if you don't have the vocabulary. Word class plays a part in that transformation. So often, things are repetitive because students can't nominalise or do any of that language…*

Similarly, Examiner F, when asked to define 'substitution', focused on the overlap between CC and LR:

> *Substitution – you mean substituting in terms of vocabulary? Like using different lexical items for the same idea, that sort of thing? I'd call that not so much coherence or cohesion as um vocabulary. (F)*

It may be that examples of lexical cohesion in longer sample scripts and the use of repetition of key nouns as a positive cohesive device with examples could usefully be included in future training materials, contrasted with some examples of scripts where repetitive vocabulary use clearly demonstrates lack of flexibility in LR. Such training would help to clarify where and in what ways to assess repetition of key vocabulary.

4.3.2 The concept of the 'essay'

Another issue worth noting was the number of times the concept of the 'essay' was invoked with reference to the assessment of coherence. Although the term 'essay' is not part of the question rubric for the Academic Writing Task 2, the assumption remains that the candidates are writing an essay and need to be familiar with this particular genre. Ten out of the 12 Phase 1 examiners referred in some form or other to the essay or particular features of this genre during the think-aloud process.

Explicit reference was made to the term 'essay' in the assessment of CC by four examiners as follows:

> *This person could be taught how to arrange this into an essay easily. (T, line 255)*

> *So they haven't really answered the question. They don't know how to write a formal essay. (D, line 560)*

> *This, look, in terms of mechanics, this essay is better than the 4 that it's getting for task fulfilment. (F, line 263)*

> *And I guess the content and organisation might be a reflection of the fact that it's only a 40 minute essay and that she's spent already 20 minutes on a different essay. (S, line 199)*

4.3.3 Overuse of cohesive devices

Another issue that caused difficulties for some examiners was the interpretation of 'overuse' of cohesive devices.

> *Um CC, I have difficulties with that because I think, do you mark someone down if they overuse, you know, techniques that they've obviously been taught because then it becomes mechanical. You know, the 'firstly', 'secondly', 'thirdly', blah blah blah. Or do you say well that's good because it adds fluency to what they're writing, you know. So that's confusing. (E)*

4.3.4 Differentiating between the band levels for CC

Some examiners in our study complained about the lack of precision in the CC descriptors. For example:

> *The paragraphing may be inadequate or missing.' I just find those 'may bes' very wishy washy. It either IS or it's not. (M)*

> *The difference between 'may be missing' and 'may be no paragraphing' and 'no paragraphing' – they're the same thing. (T)*

> *... reference and substitution. I know what it means, but I'm just thinking 'inadequate'? 'Inaccurate'? What is 'inadequate'? [...] 'inadequate' – that doesn't mean anything to me. (P)*

4.3.5 Fitting the scripts to the band descriptors

Examiners talked about the difficulty of 'fitting' the script to the band descriptors. This metaphor of 'fitting' was used by six of the 12 examiners in the Phase 1 interviews, for example:

> *[The criteria] are equally difficult to understand. I find them all difficult. Trying to fit an individual's writing into these band descriptors when bits don't fit. Which descriptor sentences you are going to go with and which can be overlooked. (L)*

> *We decide where we're going to fit them [the scripts] rather than where they necessarily fit. (A)*

One of the most experienced examiners commented that the descriptors did not always match her intuitive impression of the candidate's level.

> *Sometimes I find that becomes a bit frustrating because I feel like no, no, this one should be a 4, but when I look at the descriptors it's not. (F)*

Examiners dealt with the complexities of this issue in various ways. Evidence from the think-aloud data suggested that although examiners conscientiously tried to mark analytically, the more difficult it was to fit the script to the band descriptors, the more likely they were to fall back on an intuitive approach to marking. The following extract shows Examiner F's careful attempt to mark analytically.

175 *[reading the descriptors] The writing may be repetitive due to inadequate and or inaccurate use of reference and substitution*
176 Well, that's not the case.
177 It is repetitive – the structure
178 because it makes the same point at the beginning of each paragraph more or less.
179 But, that's not due to inadequate or inaccurate use of reference and substitution. It's due to repetition of a single idea.
(F, lines 175-179)

But later she confessed that she tends to return to marking impressionistically.

296 In the end I look at it and I think what do I think this is?
297 What mark should be given?
298 Does it actually work out to that impressionistic mark?
(F, lines 296-298)

However, Examiner D explained in the interview that profile marking against the band descriptors took precedence in the end.

> *So I'll try to, I'll try not to go too globally and I'll know that sometimes although they've come out with some mark, that could be, although it doesn't seem right, it could be just that's the way it is, because of the banding. (D)*

4.3.6 The length of the CC band descriptors

Another issue which emerged is the length of the descriptors. The CC descriptors are longer and more complex than for the other criteria, and include four sub-sections in the central bands: coherence, cohesion, referencing and substitution, and paragraphing. As Examiner K noted, trying to understand the descriptors seemed to distract her attention from the script.

> *It's quite obvious that CC has twice the words of any of the other descriptors… I'm looking at these [descriptors] 28 times and because of the way it's structured, you can't assume a familiarity with them. So […] it's taking my attention off the language of the script. (K)*

Wolfe, in his study of essay reading style and scoring proficiency, made a similar observation:

> *The process of reading an essay and using a scoring rubric to evaluate it places a heavy cognitive demand on the scorer. Scorers who are not well prepared to use a particular scoring rubric may find themselves preoccupied with trying to remember the elements of the rubric to the extent that they must break the scoring task down into smaller tasks to handle this cognitive load. (Wolfe 1997, p 102)*

The length of time devoted to reading the CC band descriptors (see Table 2) may also reflect the cognitive demand placed on examiners.

4.3.7 Interpreting the question

As discussed earlier, the amount of time devoted to the assessment of CC as compared to the time devoted to the other criteria was used as one possible indicator of the level of difficulty examiners have in assessing CC, and the number of overlaps between TR and CC was identified as a possible indicator of ambiguity in the interpretation of the band descriptors. However, it should be noted that the actual task question may also have contributed to the amount of time examiners spent on both TR and CC in Phase 1 of this study and possibly also to the overlaps between TR and CC. This factor may also have contributed to the degree of examiner hesitancy in assessing TR and CC and, therefore, needs to be borne in mind when interpreting the results of the think-aloud data.

Question A (Appendix 1) required testees to accept a statement as fact and then put forward their views as to the advantages and disadvantages of this fact. Rather than accepting the first statement as fact, a number of testees seemed to view the initial statement as an arguable proposition with which they either agreed or disagreed before discussing the advantages or disadvantages of the same.

Several examiners made some insightful comments on this issue. As Examiner A put it:

> *You know the rubric says countries are becoming [more and more similar etc] and is this a positive or negative development. It's not really asking you to disagree with the rubric. I guess if you do disagree, you'll have to say so. I don't like this task. [This is a] problem with IELTS. You can't – if you get an un-interpretable topic – these people don't know what their position is til they get to the end. (A)*

Examiner L expressed it even more clearly.

> *I think there are questions with the actual task. The statement, 'Countries are becoming more and more similar because people are able to buy the same products anywhere in the world' – um – the question doesn't actually allow you to disagree with that, whereas the last [candidate] was disagreeing with that statement so I guess they really weren't on topic, on task. This one is also disagreeing with that [statement] although is closer to – they disagree with it but they accept it and then they talk about whether it is a positive or negative development. They comment on – um – they disagree with the statement but then they move on to comment on the questions… Sometimes these questions… just the wording – sometimes the wording can be a bit difficult for the candidate and for the examiners. (L)*

Further research is necessary to explore the extent to which the wording of the questions for Academic Writing Task 2 has an impact not only on candidates' performance but also on examiners' assessment processes.

4.4 Research Question 3

To what extent do examiners differ in their marking of coherence and cohesion in Task 2 of the Academic Writing module?

When the examiners' overall scores across all four criteria (TR, CC, LR and GRA) were correlated with the standardised scores, the scores of 53 examiners were found to have correlations of 0.8 or more while the scores of two examiners had correlations of less than 0.8. However, the correlation of only one examiner was outside the 95% confidence interval of 0.68-0.88, (r=0.55) (Appendix 6).

When the scores for CC were examined, the scores of 44 examiners had reliability above the target 0.8 correlation with standardised scores, however 11 had correlations lower than 0.8, of which there were two whose correlations were significantly lower than the 0.8 target. The Spearman correlations of these examiners' scores with the standard scores were r=0.49 and 0.65 (Appendix 6).

All Levene's tests were non-significant, indicating that the different transformed variables had equal variance and therefore satisfied a major assumption needed for means analysis by ANOVA or t test. Further, all variables were also found to be normal according to the Shapiro-Wilk test. Across all criteria, there were three examiners who had mean scores which were, uncorrected for familywise error, significantly different from the mean standard scores across all four marking criteria (see Appendix 10). That there were only three out of 55 examiners who differed from the standard mean scores strongly suggests that there is no significant issue with the leniency or harshness of the markers.

Examiner	*t* Value	Significance	Mean score	Harsher or more lenient
BA13	t_{94}=1.994	p=0.049	4.92	Harsher
BAR4	t_{94}=2.436	p =0.017	6.35	More lenient
ABR12	t_{94}=2.081	p =0.04	6.19	More lenient

Table 9: Significant t tests for difference from the mean standard score of 5.54

When only CC was looked at there were no examiners whose scores were significantly harsher or more lenient than the standard scores (Appendix 11).

There was a significant difference between the strengths of correlations of all examiners against the standard scores between the four different criteria, F (3,216) =7.22, p<0.05. Bonferroni adjusted contrasts indicated that CC had the lowest mean correlation and was not significantly different to TR, but both TR and CC were significantly different to LR and GRA.

Task Response (TR)	Coherence and Cohesion (CC)	Lexical Resource (LR)	Grammatical Range and Accuracy (GRA)
0.8597	0.8511	0.8927	0.8951

Table 10: Mean correlations of all examiners with the standard scores

The presentation order of the scripts had no significant effect on the reliability of examiners' CC scores.

4.5 Research Question 4

What effect do variables such as examiners' qualifications and experience have on the marking of coherence and cohesion?

A Spearman correlation of number of years' teaching with the reliability of the examiners' scores on CC shows that there was no relationship between the number of years of full-time teaching experience and CC (r=0.091, p>0.05). The mean correlations of CC with standard scores were compared based on examiner characteristics and no significant differences in reliability based on those characteristics were found (Tables 11 and 12).

In terms of examiner qualifications, there was no difference in the reliability of CC between the 22 people who had a masters degree in either ESL or Linguistics (t_{46}=1.440, p>0.05) and the 26 who had no masters degree. Similarly, there was no difference in reliability for the 34 people who did not have a masters degree in ESL compared with the eight people who did (Z=1.025, p>0.05). Nor was there any difference on the basis of whether or not the examiners held any Linguistics qualifications (n=20) compared to the people who did not (n=35) (t_{53}=0.397, p>0.05). When the 13 people with both ESL and Linguistics qualifications were compared to the 42 people with only one or no language specific qualification, there was also no difference in the reliability of CC marking (Z=0.198, p>0.05).

Whether or not the examiners had taught academic writing more (n=36) or less frequently (n=16) than five times did not significantly affect the reliability of CC marking (t_{50}=*1.440*, p>0.05). Similarly, whether or not the examiners had marked IELTS for less than two years (n=22) or greater than five years (n= 17) had no effect on the reliability of marking CC (t_{37}=*0.726*, p>0.05). These findings accorded with Phase 1 of the study in which no differences between new examiners and experienced examiners could be established.

The Mann-Whitney U test for two independent groups found that those who marked almost every week (n= 9) were no more reliable in their marking of CC than examiners who marked less often than every two months (n= 8) (Z=0.096, p>0.05). Those who have mainly taught at lower levels (n=12) were not significantly less reliable in their marking than those who taught at upper intermediate or advanced levels (n= 42) (Z=0.458, p>0.05). Whether or not examiners prioritised flow (n=45) or structure (n=8) had no significant effect on the marking of CC (Z=0.174, p>0.05).

Comparison	N in each group	Mean Spearman correlations	Standard deviation
Those with any masters degree/ compared with those who do not	22/26	0.806/0.866	0.138/0.058
Linguistics qualifications/no linguistics qualifications	20/35	0.858/0.847	0.072/0.086
Taught academic writing more than five times/less than five times	36/16	0.867/0.819	0.068/0.139
Greater than five years of marking experience/less than two years	17/22	0.850/0.856	0.060/0.96

Table 11: Means and standard deviations of factors which were compared using t tests

Comparison (lower rank means lower reliability)	N in each group	Mean rank	Mann-Whitney U
Masters in ESL/No masters in ESL	8/34	18.06/22.31	104
Both ESL and linguistics qualifications/one qualification only	13/42	28.77/27.76	263
Mark almost every week/less often than every two months	9/8	8.89/9.12	37
Teaching upper intermediate advanced/intermediate or elementary	42/12	29.98/29.33	274
Flow prioritised/structure prioritised	45/8	26.84/27.88	187

Table 12: Results of non-parametric tests for difference

To summarise, analysis of the reliability of examiners' marking of CC in Academic Writing Task 2 showed a reasonably high degree of reliability, with no effects evident for examiner characteristics. Nevertheless, although the difference was not statistically significant from TR, it was significantly different from LR and GRA, and had the lowest mean correlation. Even though examiners are marking relatively reliably, construct validity may still be an issue given the variation in focus by a number of examiners on different features of CC and the use in their assessments of several features not explicitly noted in the band descriptors. We turn next to the final research question to explore the degree to which the training materials clarify examiners' perceptions of CC.

4.6 Research Question 5

To what extent do existing training materials clarify perceptions of CC?

The IELTS Examiner Training Materials 2004 (Writing) 2007 edition includes Powerpoint slides, 28 of which are used to explain all four criteria for the Academic Writing Task 2. Seven slides are used to provide explanations for TR, four slides for CC, ten slides for LR and seven for GRA.

The first CC slide defines coherence as pertaining to logical progression of information and argument, organisation of ideas and paragraphing, while cohesion pertains to the relationship between ideas and the 'apposite' use of cohesive devices. Trainees are informed that the CC criterion is concerned with 'the overall organisation and fluency of the message'. The remaining three slides deal with 'cohesive devices' including 'sequencers and linking devices', and 'referencing and substitution'.

Seven brief, de-contextualised examples are given on these slides to illustrate the overuse, misuse, inaccurate and inappropriate use of such devices. Coherence is explained briefly but not illustrated at this point. Trainees' attention is also drawn to the overarching statement at the beginning of each descriptor, which in the case of CC is always to do with coherence. They are also reminded to observe the bolded ceilings, particularly the ceiling at Band 5 CC, which restricts the grade for scripts which do not contain paragraphs.

Examination of the training materials reveals that there are fewer explanatory slides for CC than for the other criteria and fewer examples of key features of CC than for the other criteria. Furthermore, there are no definitions of key terms used in the band descriptors to help examiners develop a common language to talk about this criterion.

During the Phase 1 interviews, the examiners were invited to talk about how well the training materials prepared them for marking CC. Examiners generally appreciated the training but one examiner was more critical.

> *It's like you're a doctor and they said here's the guide to today's surgery and here's a scalpel. Good luck. Don't hit any of the wrong blood vessels. (K)*

All the examiners, except one, expressed a desire for greater certainty about how the CC band descriptors relate to scripts. Five of the examiners commented particularly on the value of discussion about how the band descriptors apply to particular examples; they felt that more time could have been devoted to this during the training. Another high priority for the interviewees, especially the new examiners, was getting feedback after the training: three examiners could not recall receiving any feedback on their marking. Three of the new examiners suggested that a buddy system or some form of mentoring would be valuable in this regard. As Examiner L remarked:

> *I've been examining for just over 12 months now so I don't feel like I've gained experience [as an IELTS examiner]. You know because you're not getting feedback… to know whether I'm on track. I'd like to get a bit more feedback especially initially to know whether I'm interpreting things [correctly]… I recently had a review of my marking and there were comments but I don't have the text, the writing to relate it to. (L)*

In the Phase 2 survey, we asked examiners to rate the examiner training or last standardisation in the assessment of coherence and cohesion on a five-point Likert scale from poor to excellent. The majority of examiners rated their last training in relatively positive terms as average or above:

- 47% (n=26) gave a rating of 'average'
- 35% (n=19) gave a rating of 'very good'
- two examiners gave a rating of 'excellent'
- 15% (n=8) of examiners gave a rating of 'below average'.

This response is not as positive as McDowell's (2000) evaluation of IELTS examiner training. In particular, McDowell's respondents expressed appreciation for the 'homework scripts' which they could take home to consider – a resource not included in the current examiner training procedure, and one which several participants mentioned they would like.

Examiners were also asked to indicate how much they remembered about their training or last standardisation in relation to CC. While almost a quarter (24%, n=13) indicated that they remembered a great deal, just over a quarter (27%, n=15) indicated that they remembered 'not much' and 45% (n=25) remembered only 'some' of their training in relation to CC. The findings for this question appear to be in keeping with comments made by several examiners interviewed in the qualitative phase of this study, who expressed some frustration at how much they were expected to remember without anything to take away with them. As Examiner L put it:

> *In lots of ways you don't have the opportunity to reflect because one – you can't take the descriptors away. I mean, I felt like I would have liked to have taken the descriptors away to work on having worked through some of the writing, to work on what I thought was the most important parts of it. (L)*

Examiners were asked if they were aware that at each testing centre they can access standardised scripts for regular revision. While 78% of examiners indicated that they were aware of this fact, 12 examiners (22%) or almost a quarter did not know they had access to such scripts for revision purposes.

Another useful guide for revision purposes is the 'Information for Writing Examiners' booklet. This contains the rating scales and a summary of the key features to assess for each criterion. One Phase 1 examiner indicated how important she found these guidelines as a means of revision and that she read them every time she examined in order to put her 'in the zone' for marking scripts. Phase 2 examiners were also asked how frequently they read this booklet before marking. As can be seen in Table 13, the findings suggest that a significant number of examiners do not follow this procedure, with 25% (n=14) stating they rarely read the booklet and 14% (n=8) that they never read the booklet containing the definitions of the features to assess for each criterion.

SCALE	n	%
Every time I mark	11	20%
Often	12	22%
Sometimes	10	18%
Rarely	14	25%
Never	8	14%

Table 13: Frequency with which examiners read the writing examiners' instruction booklet before marking

Despite the finding that a significant proportion of examiners do not recall much about their training in the assessment of coherence and cohesion, responses for the above question indicate that at least some examiners may not make the best use of materials available to them for revision purposes. It may be that examiners need to be reminded of these materials and the usefulness of particular revision strategies on a more systematic and regular basis.

A number of suggestions were made by the Phase 1 examiners for improving examiner training with particular reference to the assessment of CC. The suggestions were as follows:

1. A detailed analysis of coherence and cohesion in one or two full length scripts for each band level showing all cohesive devices and illustrating for example, their misuse, overuse or omission.
2. Revision of the CC band descriptors to ensure greater consistency in terminology.
3. A glossary of key terms for describing coherence and cohesion with definitions and examples to be included in the instructions for writing examiners.
4. More mentoring and feedback in the first year of examining.
5. An online question and answer service available for examiners.
6. Use of colours to pick out the key features across the bands in the rating scale.
7. A list of dos and don'ts for marking scripts (eg don't compare assessments of one script against another) to be included in the examiners' instructions booklet.
8. A step-by-step guide to the recommended process (or processes) to follow (eg refresh your memory of the band descriptors before marking) to be included in the instructions for writing examiners' booklet.
9. Online training materials with exercises for revision and reflection.

The Phase 2 examiners were asked to indicate their level of agreement with each of these suggestions from 1 (strongly agree) to 5 (strongly disagree). The results show that examiners would particularly appreciate: the provision of a glossary of key terms in CC (95% agreement); online training materials and revision exercises (90% agreement); a detailed analysis of CC in a number of scripts at each band level (89% agreement); and the revision of the band descriptors for CC to ensure greater consistency of terminology (77% agreement) (see Table 14).

Although no examiners indicated disagreement with the suggestion for a detailed analysis of coherence and cohesion in a number of essays, a small percentage of examiners disagreed with each of the other suggestions, possibly because they personally did not feel the need for them, but in each case the majority of examiners supported the idea of the recommended improvements.

	SUGGESTIONS	**Strongly agree**		**Agree**		**Neither agree nor disagree**		**Disagree**		**Strongly disagree**	
		n	%	n	%	n	%	n	%	n	%
1	A detailed analysis of CC in 1 or 2 essays	**29**	**53%**	**20**	**36%**	6	11%	0	0	0	0
2	Revision of CC band descriptors	18	33%	24	44%	11	20%	2	4%	0	0
3	A glossary of key terms with definitions and examples	**28**	**51%**	**24**	**44%**	2	4%	0	0	1	2%
4	More 1st year mentoring	11	20%	32	58%	10	18%	2	4%	0	0
5	Online Q & A service	12	22%	25	45%	15	27%	3	5%	0	0
6	Use of colours to identify features across bands	11	20%	15	27%	23	42%	6	11%	0	0
7	A list of dos and don'ts in instructions booklet	6	11%	26	47%	21	38%	2	4%	0	0
8	A step-by-step process guide included in instructions booklet	7	13%	30	55%	15	27%	3	5%	0	0
9	Online training materials with revision exercises	**30**	**55%**	**19**	**35%**	5	9%	0	0	1	2%

Table 14: Levels of agreement about suggestions for changes to the training

Examiners were then asked which three suggestions for revision of the training they considered the most important by marking 1, 2 or 3 beside each chosen item in the list. In terms of the highest number of votes for first, second and third preferences, the data indicates that a detailed analysis of CC in some sample scripts was ranked first by 50% (n=28) of examiners; a glossary of key terms by 35% (n=19) of examiners; and revision of the band descriptors for coherence and cohesion by 22% (n=12) of the respondents. Additional examiner comments can be seen in Appendix 11. The need for more training in the assessment of CC seems to be a common theme running through all these comments, with requests made for more examples, clearer explanations of key terms, more practice and more opportunity to ask for clarification from experts.

5 SUMMARY OF RESULTS

At the start of this project, we raised a number of research questions as follows:

1. Do examiners find the marking of CC more difficult than the marking of the other three criteria
2. What are examiners looking for in marking CC in Task 2? What features of Task 2 texts affect their decision-making in relation to the CC band descriptors?
3. If examiners are interpreting the band descriptors in slightly different ways, are they marking reliably?
4. What effect do variables such as examiners' qualifications and experience have on their marking of coherence and cohesion?
5. To what extent do existing training materials clarify examiner perceptions of coherence and cohesion?

Both the qualitative Phase 1 and the quantitative Phase 2 data in this study provide some important insights in relation to each of these questions, which are summarised below.

5.1 Question 1

Data from the Phase 1 think-aloud protocols, the interviews and the Phase 2 surveys indicated that the majority of examiners in this study find the assessment of CC more difficult than the marking of the other three criteria and are less confident when marking this criterion. These findings are in line with those of Shaw and Falvey (2008, p 165). In the think-aloud data, examiners spent more time on the assessment of CC and TR than on LR and GRA. They took longer reading the CC band descriptors and hesitated slightly more when assessing CC as compared to the other criteria.

5.2 Question 2

Overall, examiners in the think-aloud protocols paid much more attention to the assessment of 'coherence' than to the assessment of 'cohesion'. Almost three-quarters of the segments on the assessment of CC were devoted to examiners' evaluation of coherence and slightly more than a quarter to the assessment of cohesion.

However, closer examination of the segments in Phase 1 focusing specifically on 'coherence' indicated that approximately one-third of these were devoted to gaining an overall impression of the text, involving subjective interpretations of the scripts. Slightly more than a third were focused on examiners' assessment of aspects of logic, logical organisation, logical progression or the logical sequencing of ideas and the remaining third were focused on the assessment of paragraphing.

Of those segments dedicated to the assessment of cohesion, the majority were focused on examiners' assessment of 'cohesive devices', 'coordinators', 'discourse markers' and 'linking words' – terms examiners seemed to use interchangeably. Comparatively few segments were focused on examiners' assessments of 'reference' and/or 'substitution'.

There was, however, variation between the 12 examiners in the degree to which they focused on the assessment of features of coherence as opposed to the assessment of cohesion (see Table 6). For example, at one end of the scale, Examiner K focused on features of coherence in 39% of her think-aloud transcript and 61% on features of cohesion. At the other end of the scale, 90% of the segments of Examiner A made explicit reference to features of coherence and only 10% to aspects of cohesion.

The other examiners ranged between these two examiners in the degree to which they referred to either coherence or cohesion.

Differences in the interpretation of the importance of certain features in the CC band descriptors seemed to be evident in examiners' assessment of reference and substitution. While eight examiners in the think-aloud protocols assessed instances of reference and substitution in the set of 10 scripts, four examiners did not appear to assess these features in the same set of scripts.

Examiners used the terminology of the CC band descriptors extensively. However, a number of other terms were used quite frequently, including the terms, 'overall structure' used synonymously with logical organisation; 'linking words' in place of discourse markers and similar terms; and 'flow'. There were indications that a few of the examiners were using this in place of 'logical progression', but others were using this term in a very general way. As one examiner put it: 'it's a gut feeling'.

Additional terms explicitly referred by a number of examiners in the think-aloud protocols that are not in the band descriptors, included: 'introduction', 'conclusion', 'essay', 'argument' and 'topic sentence'. These examiners appeared to be looking for a well-structured essay, featuring the classical introduction–body–conclusion format and clearly demarcated paragraphs with topic sentences, despite the fact that the task rubric or prompt wording does not instruct candidates to use this format.

Although paragraphing plays an essential role in guiding most examiners' marking, some examiners found it difficult to reconcile their marking of coherence with the paragraph ceiling specified in Band 5. Several examiners also noted that it was difficult to differentiate between the 'relative terms' in relation to paragraphing between some of the bands.

The definitions examiners provided for the key terms from the band descriptors for CC indicated that although the majority of examiners provided similar definitions for 'coherence', 'cohesion', 'cohesive devices', 'reference' and 'substitution' in broad terms, some examiners were less able to define these terms, indicating perhaps that they were unclear about the meanings of these terms.

A small number of overlaps between the assessment of TR and CC in the think-aloud protocols appeared to relate to the fact that some examiners had difficulty distinguishing between particular features of the two band descriptors, especially in relation to the assessment of:

- how clearly a position is presented and supported (TR)
- how clearly the message is logically organised (CC)
- the development of main ideas (TR)
- the logical progression of ideas (CC).

There was also some question over the marking of 'relevance' and whether this should be assessed under TR or CC. Some examiners indicated uncertainty about whether to assess lexical cohesion, the use of synonyms and the repetition of key nouns under CC or LR.

Although differences in marking style were not the focus of this study, it was noted that three of the more experienced Phase 1 examiners appeared to mark CC more intuitively or holistically with an emphasis on the general flow or fluency of the text, while four of the six less experienced examiners appeared to be more analytical, referring in a systematic way to most or all of the features listed in the band descriptors.

5.3 Question 3

Despite the fact that some examiners in the think-aloud protocols appeared to be interpreting the band descriptors in slightly different ways, and some examiners in Phase 2 seemed to be uncertain about the meaning of some linguistic terms such as 'substitution', they were nevertheless marking very reliably. In Phase 2, 55 examiners marked 12 scripts across four criteria making 2640 observations in all. Correlations of these scores against the standardised scores across all four criteria indicated a high degree of reliability with all but one examiner, with overall correlations above 0.8 and eight of these with correlations above 0.9. The marking of CC was slightly less reliable than the marking of the other three criteria but still remained within acceptable levels.

5.4 Question 4

No significant effects could be found for any examiner characteristics in this study. No effect could be found for IELTS marking experience, higher qualifications, training in linguistics, and either the level of most teaching experience or for the number of years of teaching experience. This would seem to suggest that the IELTS training, certification and re-certification processes have been effective in ensuring the reliability of examiners regardless of differences in their background and experience.

5.5 Question 5

Examiner training materials did not appear to provide as much advice about the assessment of CC as for the other three criteria. Nevertheless, most examiners were reasonably satisfied with the training they had received in relation to Coherence and Cohesion, although a number felt that more time was needed for discussion and reflection. A number of examiners noted that they did not recall a great deal about their training in CC but there were indications that a significant number of examiners rarely read the 'Information for Writing Examiners' booklet and some examiners were unaware that they could access standardised scripts for revision purposes.

6 DISCUSSION AND RECOMMENDATIONS

Findings from this study show that there was greater emphasis in the think-aloud protocols on the assessment of coherence over the marking of cohesion and a degree of variability in the attention examiners paid to different features of the CC band descriptors, which suggests that examiners may differ in their interpretations of these features. In addition, a number of examiners appeared to have an incomplete understanding of some of the linguistic terms used in the CC band descriptors. While examiners are marking to a high standard of reliability, these findings nevertheless have implications for the construct validity of the test, which it may be possible to address in the following ways.

1. Additions or refinements to examiner training for CC.
2. A possible re-assessment of and fine tuning of the band descriptors for CC.
3. Consideration of the task rubric so that candidates unfamiliar with the essay genre are not disadvantaged.
4. Further discourse studies of aspects of coherence and cohesion in sample texts at different levels.

6.1 Suggested additions or refinements to examiner training for CC

It is generally agreed that the revised rating scales introduced in 2005 as a result of the four-year long IELTS Writing Assessment Revision Project are a great improvement on the previous scale. The introduction of four analytic band scales for four different criteria and greater precision in the wording at each band level led to a much better testing instrument which, together with the introduction of the re-training, certification and re-certification process, has resulted in improved examiner reliability (Shaw and Falvey, 2008).

Nevertheless, as Allison (1999, p 180) points out, analytic marking itself may still be impressionistic. The marking of coherence, in particular, depends largely on the reader's subjective perception of the text, as writers such as Jones (2007) contend. The think-aloud protocols would seem to confirm this perception, in that approximately a third of coherence segments were focused on examiners gaining a general overall impression of the clarity, flow or coherence of the text.

While subjective marking of propositional coherence cannot be entirely avoided, it may be helpful to try to minimise its impact. While study of exemplar scripts are an important part of the current training, more such scripts at different levels, but with all cohesive ties and logical connections clearly annotated, could be usefully included in the training materials. As the working group of the IELTS Writing Assessment Revision Project agreed in Phase 2 of that project, '… training materials should be replete with contextualised examples…' (Shaw and Falvey, 2008, p 44). Study of such annotated scripts would allow examiners to discuss the different ways in which propositional coherence can be identified in text and share their understandings.

It may also be that the study of thematic progression in sample texts could greatly aid examiners to appreciate more fully the ways in which ideas can be logically connected. As described in the literature review, Knoch (2007) had some success in implementing a scoring system based on Topic Structure Analysis (TSA), which analyses thematic progression. Although examiners found it hard to learn how to apply the system, they achieved higher reliability than when using the former five trait scoring descriptors. The identification of coherence breaks, as discussed by Knoch, could provide a useful analytical tool for examiners in assessing 'flow' or 'logical progression'. While such detailed study may not be practical as part of the existing training, the development or re-introduction of a homework pack containing sample texts and exercises related to CC might be made available for those examiners who may need or want to improve their knowledge of this important area.

The findings of this study indicate that a number of examiners may not fully understand some of the linguistic terms used in the band descriptors for CC. Indeed, it is possible that some examiners may never have come across some of the terms used in the CC band descriptors before becoming examiners and could be embarrassed to admit ignorance about the terminology used. As Shaw and Falvey (2008) noted, '... the concept of *coherence*, when linked to *cohesion,* creates difficulties for examiners who, perhaps, are not familiar with the recent literature in text linguistics' (2008, p 174). The working group of the IELTS Writing Assessment Revision Project also agreed that, '... the terms 'reference' and 'substitution' should be fully explained to examiners' (p 44). The inclusion of a glossary of key terms in the training materials as recommended by one of the working group (Shaw and Falvey, 2008, p 158), together with more detailed explanations by trainers, would provide examiners with more opportunity to clarify the meaning of these different concepts and linguistic terms used in the CC band descriptors.

The working group also agreed that, '... examiners must be fully sensitised to the 'grey' areas' (Shaw and Falvey, 2008, p 44). Among such areas identified in this study are those relating to the small number of overlaps between TR and CC, where some examiners did not seem to be certain about which criteria they should be using to assess particular features of the text. Future refinements to the training material may need, for example, to clarify the difference between the assessment of the development of ideas (rated under TR) and the assessment of the logical progression of ideas (rated under CC). Similarly, training materials may need to illustrate the difference between the degree to which a position is presented (currently assessed under TR), and the degree to which the message/ideas are conveyed (assessed under CC).

Other grey areas which may need further explanation and exemplification in the examiner training include whether 'relevance' should be assessed under CC or TR and whether synonymy, lexical chains and the repetition of key nouns should be assessed under CC as part of lexical cohesion, or under LR as a indication of the flexibility or otherwise of vocabulary use.

The interviews also highlighted the need for more feedback and more mentoring for new examiners. Perhaps examiners who lack confidence, in particular, could request some extra mentoring sessions with a senior examiner. Weigle (2002, p 131) recommends that examiners be informed about the amount of variability in scoring that is acceptable and be reassured that they are not always expected to be 100% accurate. Such reassurance might be valuable for examiners with low confidence levels. A background reading list on coherence and cohesion might also be appreciated. Examiners could also be reminded to take full advantage of the revision materials already available to them.

6.2 Possible re-assessment of and fine tuning of the band descriptors for CC

The study also points to the possible need to fine tune some of the wording in the band descriptors for CC. Not only are the descriptors longer and more complex than those for other criteria, but the small number of overlaps between CC and TR and, to a lesser extent, between CC and LR discussed in the previous section may have the potential to undermine the construct validity of the criterion. As well as addressing these issues in the training, fine tuning may help to reduce the number of 'grey areas' that examiners noted in the IELTS Writing Assessment Revision Project (Shaw and Falvey, 2008, p 44). For example, it could be argued that 'relevance' could be better assessed under TR throughout the bands rather than being included under CC at Band 2.

Another aspect of the band descriptors that may need to be re-considered is the ceiling for paragraphing in Band 5. Although examiners 'followed the rules' in this regard, the data showed that some were uncomfortable doing so. If the essay genre is expected, then it is reasonable to demand appropriate paragraphing. However, to impose this ceiling may disadvantage candidates who can write coherently, but who may not be aware of the cultural requirements of paragraphing in academic English. This point was made several times by participants in the IELTS Writing Assessment Revision Project (Shaw and Falvey, 2008), who reported that the paragraphing criteria were 'somewhat strict' (p 246) and suggested that candidates who had attended IELTS preparation classes would be advantaged. As one senior examiner noted: '... focus on conventional paragraphing gives undue advantage to the learned/practised responses we get such a lot of...' (p 49).

Another issue of interest is the role played by lexical cohesion, which Hoey (1991, p 9) claims is the most important form of cohesive tie. He draws our attention to the fact that in the analysis of cohesion in seven different types of text conducted by Halliday and Hasan (1976), lexical cohesion accounted for almost half of all cohesive ties. If that is generally the case, it might be possible to place greater emphasis on the assessment of lexical cohesion in the Academic Writing Task 2. It may be useful to consider this issue as part of future revisions of the band descriptors.

6.3 Revision of the task rubric to minimise candidate disadvantage

As part of the IELTS Writing Assessment Revision Project (Shaw and Falvey, 2008), revisions were made to the task rubric for Task 1 and Task 2 to omit reference to specific genres such as 'report' or 'essay', together with the words 'argument' and 'case', on the grounds that candidates would not necessarily be familiar with the requirements of these genres. Nevertheless, the Academic Task 2 still asks candidates to present a point of view or discuss the advantages or disadvantages of a given statement. In the minds of the examiners, such answers call for an introduction, well-formed body paragraphs and a conclusion in which the answer to the question is clearly stated: the typical essay genre (Connor 1990; Mickan and Slater 2003).

If the term 'essay' is avoided in the task rubric but examiners are nevertheless marking for features of the essay genre, this raises the question of whether some candidates may, in fact, be disadvantaged rather than the reverse by the omission from the task wording of explicit instructions about the genre type expected. As Weigle (2002, p 63) points out, the wording of prompts including whether or not the 'rhetorical task or pattern of exposition is explicitly stated' can have an impact on candidates' performance.

6.4 Further studies of aspects of coherence and cohesion in sample texts at different levels

As noted in the literature, there have been calls for scale writing to be based on empirical studies of language use rather than the judgments of experts, in order to increase the validity of such tests (Fulcher 1987; North and Schneider 1998; Turner and Upshur 2002). Further studies of coherence and cohesion in text, using a corpus or discourse analysis approach, such as those of Kennedy and Thorp (2007) as well as Mayor, Hewings, North, Swan and Coffin (2007), could provide important new reference sources for future revisions of the band descriptors. Further studies of this kind would also provide more examples of the analysis of CC at different band levels. Extracts from these studies could usefully be employed in future training sessions to illustrate more precisely particular features of CC at the different levels.

7 CONCLUSION

In many respects this study of examiner marking of CC in IELTS Academic Writing Task 2 presents an encouraging picture. Although the marking of CC was less reliable than for LR and GRA, the overall reliability of CC was found to be acceptable. In addition, examiners' qualifications and experience were found to have no significant effect on the marking of CC, possibly indicating the effectiveness of current training procedures for maintaining examiner reliability. Examiners worked very professionally, adhered closely to the band descriptors, and were keen to internalise the wording of the band descriptors. As Examiner P put it, 'I just want to carry the words [of the descriptors] in my head'.

As Lumley (2002, 2005) found, examiners are engaged in a highly complex process in their marking, involving an intuitive appreciation of text. However, it is possible that this very intuition poses a threat to the construct validity of the CC criterion, as examiners may differ in their interpretation of the descriptors. Nevertheless, we argue on the basis of this study that some measures could be taken to improve the assessment of coherence and cohesion including some additions and refinements to the examiner training materials, possible fine tuning of the wording of the band descriptors for CC, and a further revision of the task rubric or wording of the prompt. Such measures have the potential to improve the construct validity of the test, and could also increase examiners' confidence in their marking of this criterion, as well as improve the degree to which examiners share understandings of coherence and cohesion so that they can participate as a more integral part of a shared community of practice.

ACKNOWLEDGEMENTS

We would like to thank the following people for their invaluable support and assistance: IELTS Australia, especially Jenny Osborne, former Regional Manager; Cambridge ESOL for providing scripts and standardised scores for this study; the IELTS administrators who facilitated the study in three different Australian cities, and all the examiners who willingly volunteered and participated with good humour in the study.

Special thanks to Dr Grenville Rose for providing the statistical input and the report on examiner reliability and the impact of examiner qualifications and experience. Thanks also to Dr David Pedersen, statistical consultant at the University of Canberra.

REFERENCES

Alderson, J, Clapham, C, and Wall, D, 1995, *Language test construction and evaluation,* Cambridge University Press, Cambridge

Allison, D, 1999, *Language testing and evaluation,* Singapore University Press, Singapore

Barkaoui, K, 2007, 'Rating Scale impact on EFL essay marking: a mixed-method study', *Assessing Writing,* vol 12, pp 86-107

Brown, A 2000, Legibility and the rating of second language writing: an investigation into the rating of handwritten and word-processed IELTS task two essays, in *IELTS Research Reports Volume 3*, ed R Tulloh, IELTS Australia Pty Limited, Canberra, pp 131-151

Canagarajah, AS, 2002, *Critical academic writing and multilingual students*, University of Michigan Press, Ann Arbor

Canale, M, 1983, 'From communicative competence to language pedagogy', in *Language and communication*, eds J Richards and J Schmidt, Longman, London, pp 2-27

Canale, M, 1984, 'A communicative approach to language proficiency assessment in a minority setting', in *Communicative competence approaches to language proficiency assessment: research and application,* ed C Rivera, Multilingual Matters, Clevedon, pp 107-122

Canale, M, and Swain, M, 1980, 'Theoretical bases of communicative approaches to second language teaching and testing', *Applied Linguistics,* vol 1(1), pp 1-47

Connor, U, 1990, 'Linguistic/ rhetorical measures for international persuasive student writing', *Research in the Teaching of English,* vol 24 (1), pp 67-87

Connor, U, and Farmer, F, 1990, 'The teaching of topical structure analysis as a revision strategy for ESL writers', in *Second language writing: research insights for the classroom*, ed B Kroll, Cambridge University Press, Cambridge

Cox, K, and Hill, D, 2004, *EAP Now,* Pearson Education Australia, Frenchs Forest, NSW

Crow, B, 1983, 'Topic shifts in couples' conversations', in *Conversation Coherence: Form, Structure and Strategy,* eds. RC Craig and K Tracy, Sage Publications, Beverley Hills, CA, pp 136-156

Cumming, A, Kantor, R, and Powers, DE, 2001, *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: an investigation into raters' decision making and development of a preliminary analytic framework*, Educational Testing Service, Princeton, NJ

DeRemer, ML, 1998, 'Writing assessment: raters' elaboration of the rating task', *Assessing Writing,* vol 5 (1), pp 7-29

Eckes, T, 2008, 'Rater types in writing performance assessments: A classification approach to rater variability', *Language Testing,* vol 25 (2), pp 155-185

Ericsson, KA, and Simon, HA, 1984, *Protocol analysis*, MIT Press, Cambridge, MA

Faerch, C, and Kasper, G, 1987, 'From process to product: Introspective methods in second language research', in *Introspection in second language learning,* Multilingual Matters, Clevedon

Falvey, P, and Shaw, S, 2006, 'IELTS writing: revising assessment criteria and scales (Phase 5), *Research Notes* 23, pp 7-12

Fulcher, G, 1987, Tests of oral performance: the need for a data-based criteria. *English Language Teaching Journal* 42 (4), 287-91

Furneaux, C, and Rignall, M, 2007, 'The effect of standardization-training on rater judgements for the IELTS Writing Module', in *Studies in Language Testing 19, IELTS collected papers, Research in speaking and writing assessment,* Cambridge University Press, Cambridge, pp 422-445

Green, A, 1998, *Verbal protocol analysis in language testing research: a handbook,* Studies in Language Testing 5, Cambridge University Press

Halliday, MAK, and Hasan, R, 1976, *Cohesion in English*, Longman, Harlow

Halliday, MAK, and Matthiessen, C, 2004, *An introduction to functional grammar*, Arnold, London

Hamp-Lyons, L, 1991, 'Scoring procedures for ESL contexts', in *Assessing second language writing in academic contexts,* ed L Hamp-Lyons, Ablex, Norwood, NJ, pp 241-276

Hamp-Lyons, L, 2007, 'Worrying about rating', *Assessing Writing* vol 12, pp 1-9

Hawkey, R, 2001, 'Towards a common scale to describe ESL writing performance', *Research Notes* 5, pp 9-14

Hoey, M, 1991, *Patterns of lexis in texts*, Oxford University Press, Oxford

Howell, D, 1982, *Statistical methods for psychology*, PWS Kent, Boston, MA

Jones, J, 2007, 'Losing and finding coherence in academic writing', *University of Sydney Papers in TESOL 2* (2), pp 125-148

Kennedy, C, and Thorp, D, 2007, 'A corpus based investigation of linguistic responses to an IELTS academic writing task', in *Studies in Language Testing 19: IELTS collected papers; Research in speaking and writing assessment*, eds L Taylor and P Falvey, Cambridge University Press, Cambridge, pp 316-378

Knoch, U, 2007, 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence', *Assessing Writing* vol 12, pp 108-128

Knoch, U, Read, J, and von Randow, J, 2007, 'Re-training writing raters online: how does it compare with face-to-face training?', *Assessing Writing,* vol 12 (1), pp 26-43

Lumley, T, 2002, 'Assessment criteria in a large-scale writing test: what do they really mean to raters?' *Language Testing,* vol 19 (3), pp 246-276

Lumley, T, 2005, *Assessing second language writing: the rater's perspective*, P Lang, New York

Mann, W, and Thompson, S, 1989, *Rhetorical structure theory: a theory of text organization.* Information Sciences Institute, University of Southern California, Los Angeles.

Mayor, B, Hewings, A, North, S, Swann, J, and Coffin, C, 2007, 'A linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2' in *Studies in Language Testing 19: IELTS collected papers; Research in speaking and writing assessment*, eds L Taylor and P Falvey, Cambridge University Press, Cambridge, pp 250-314

McDowell, C, 2000, Monitoring IELTS examiner training effectiveness, in *IELTS Research Reports Volume 3*, ed R Tulloh, IELTS Australia Pty Limited, Canberra, pp 109-141

McNamara, T, 1996, *Measuring second language performance*, Longman, London

Mickan, P, and Slater, S, 2003, Text analysis and the assessment of Academic Writing, in *IELTS Research Reports Volume 4,* ed R Tulloh, IELTS Australia Pty Limited, Canberra, pp 59-88

Milanovic, M, Saville, N, and Shuhong, S, 1996, 'A study of the decision-making behaviour of composition markers', *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem,* Cambridge University Press

North, B, and Schneider, G, 1998, 'Scaling descriptors for language proficiency scales.' *Language Testing* vol 15 (2), pp 217–263

Oshima, A, and Hogue, A, 2006, *Writing academic English,* 4th edn, Pearson Longman, White Plains, NY

Padron, YN, and Waxman, HC, 1988, 'The effect of ESL students' perceptions of their cognitive strategies on reading achievement', *TESOL Quarterly* vol 22, pp 146-150

Paltridge, B, 2001, *Genre and the language learning classroom*, University of Michigan Press, Ann Arbor

Schaefer, E, 2008, 'Rater bias patterns in an EFL writing assessment', *Language Testing* vol 25 (4), pp 465-493

Schneider, M, and Connor, U, 1990, 'Analyzing topical structure in ESL essays', *Studies in Second Language Acquisition* vol 12, pp 411-427

Shaw, S, 2002, 'The effect of training and standardisation on rater judgement and inter-rater reliability' *Research Notes* vol 8, pp 13-17

Shaw, S, 2004, 'IELTS Writing: revising assessment criteria and scales (Phase 3)' *Research Notes* vol 16, pp 3-7

Shaw, S, 2006, 'IELTS Writing: revising assessment criteria and scales (Phase 5)', *Research Notes*, vol 26, pp 7-12

Shaw, S and Falvey P, 2008, 'The IELTS Writing Assessment Revision Project: Towards a revised rating scale', *Research Reports*, vol 1, January 2008

Turner, CE, and Upshur, JA, 2002, 'Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores', *TESOL Quarterly* vol 36 (1), pp 49-70

Watson Todd, R, 1998, 'Topic-based analysis of classroom discourse', *System,* vol 26, pp 303-318

Watson Todd, R, Khongput, S, and Darasawang, P, 2007, 'Coherence, cohesion and comments on students' academic essays', *Assessing Writing* vol 12, pp 10-25

Watson Todd, R, Thienpermpool, P, and Keyuravong, S, 2004, 'Measuring the coherence of writing using topic-based analysis', *Assessing Writing* vol 9, pp 85-104

Weigle, SC, 1994, 'Effects of training on raters of ESL compositions', *Language Testing* vol 11 (2), pp 197-223

Weigle, SC, 1998, 'Using FACETS to model rater training effects'. *Language Testing,* vol 15 (2), pp 263-287

Weigle, SC, 2002, *Assessing writing*, Cambridge University Press, Cambridge

Wolfe, EW, 1997, 'The relationship between essay reading style and scoring proficiency in a psychometric scoring system', *Assessing Writing* vol 4 (1), pp 83-106

Wolfe, EW, Kao, C-W, and Ranney, M, 1998, 'Cognitive differences in proficient and non proficient essay scorers', *Written Communication* vol 15, pp 465-492

APPENDIX 1: WRITING TASKS

WRITING TASK A

You should spend about 40 minutes on this task.

Write about the following topic:

Countries are becoming more and more similar because people are able to buy the same products anywhere in the world.

Do you think this is a positive or negative development?

Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Write at least 250 words.

WRITING TASK B

You should spend about 40 minutes on this task.

Write about the following topic:

Some people think that money is the only reason for working. Others, however, believe that there are more important aspects to a job than the salary.

Discuss these views and give your own opinion.

Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Write at least 250 words.

APPENDIX 2: SEMI-GUIDED INTERVIEW SCHEDULE (PHASE 1)

ASSESSMENT OF COHERENCE AND COHESION IN WRITING TASK 2

SEMI-GUIDED INTERVIEWS WITH SAMPLE OF IELTS EXAMINERS

Examiner's name…….. Date……..

Preamble: Thank you so much for marking the scripts. Now that you have finished, we would like to interview you. There is anecdotal evidence that some examiners still find aspects of marking the scripts particularly difficult to assess, even if they are very experienced. We'd like know whether you have difficulties with the same issues. So we are interested to hear your views on the different criteria and the procedures you follow when marking the scripts.

1) BAND DESCRIPTORS

Let's start by looking at the actual band descriptors (Get it out and look at it together)

1. Are all the band descriptors equally easy to understand?
2. If not, which ones do you find more difficult or easier to understand?

2) ASSESSING THE SCRIPTS

When you actually start marking the scripts:

3. In general, which of the four criteria do you find the most difficult and which the easiest to assess and **why?**
4. Do you sub-vocalise (or mutter quietly out aloud) when you are marking the scripts?

   If so, in what ways do you think it helps you mark the scripts?
5. Do you ever refer to the guidelines in the examiners' writing instruction booklet? If so, when and how often do you refer to them? For instance, do you read them before you begin marking?

3) ASSESSMENT OF COHERENCE AND COHESION

For our study we are looking particularly at coherence and cohesion, so we want to ask you some more detailed questions about this particular criterion

6. What do you think is the difference between coherence and cohesion?
7. What do you look for when you are marking for coherence and cohesion? (Do you separate them out when you are marking?)
8. How difficult do you find it to mark CC compared to the other criteria?
9. The band descriptors refer to a number of measures for coherence and cohesion; including 1) logical order/ sequencing 2) cohesive devices, 3) reference, 4) substitution and 5) paragraphing. How easy is it to understand these different terms? What do these terms mean to you?

10. Which of these measures you do find the most useful in evaluating CC? (Are there any which you tend not to use when assessing scripts?)

11. How confident do you feel in the marking of coherence and cohesion as compared to marking the other criteria? If you are not confident, what would make it easier for you to mark CC?

12. How much do you think your level of confidence in marking CC relates to any of the following:

	not at all	a little bit	a great deal
a) your level of experience in teaching writing	○	○	○
b) your level of experience as an IELTS examiner	○	○	○
c) the IELTS training	○	○	○
d) the clarity of the band descriptors	○	○	○

THE SAMPLE SCRIPTS

Now I'd like us to have a closer look at the different scripts you have just marked. Let's look at the ones you have just talked about on the tape and see what the standardised scores were.

13. Can we look again at script x (y and z). The scores were … To reflect on your own marking, which of the measures of coherence and cohesion eg: 1)logical order/sequencing 2) cohesive devices 3) reference 4) substitution and 5) paragraphing seemed to help you most in your assessment of this script?

TRAINING

We would also like to ask you some questions about the IELTS training:

14. What IELTS training did you receive in assessing coherence and cohesion?

15. How adequate do you think that training was?

16. Have you any suggestions as to how the training might be improved in order to help you with the marking of CC?

APPENDIX 3: MAIN CODES USED IN THE THINK-ALOUD DATA ANALYSIS

1. GB= General behaviors:

a. M=management strategies
b. R=reading
c. I= interpretation
d. J= judgement

2. SB= specific behaviors:

a. M = Management
   i. DIR = indicating direction in which they are going (Now moving on to the next script…)
   ii. EXPL = Explanation. (eg: I usually start by making an overall judgment)
   iii. CONCL = commenting on what has gone before
b. R = Reading
   i. RS = Reading scripts
   ii. RC = reading criteria
   iii. RQ = reading/referring to question
c. I = Interpretation
   i. Q = Questioning (Is that meant to be x …?)
   ii. RE = Rephrasing testees' position (This person seems to be arguing…)
   iii. IQ = Interpreting the question (The question seems to require…)
   iv. H = Hedging (I think it's a… I'll probably give it a…)
d. J = Judgement
   i. EVAL = evaluating the writing in general terms. (It's pretty good overall)
   ii. GR = GRADING (That's a 6. I'll give it a 7. etc.)
   iii. JU = Grading Justification (…because the paragraphs don't link)
   iv. HES = Hesitancy/uncertainty in decision making (it could be higher/ lower!)
   v. PERSON = Personifying/personalising the writer
   vi. RP = Responding to the writer's position (This candidate has an axe to grind!)
   vii. INT = intuition

3. The focus of attention/ items being referred to:

a. TR = task response
b. CC = coherence and cohesion
c. LR = lexical resources
d. GRA = grammatical range and accuracy
e. ALL = all criteria/whole
f. SELF = examiner
g. TXT = the text being assessed

4. Coherence codes:

a. Coherence = Coherence
b. Meaning/message = M
c. Argument = ARG
d. Logic/al =LOG
e. Logical organization = LOG ORG
f. Logical progression = LOG PRO
g. Relationship of ideas = REL
h. Fluency/flow? = FL
i. Clarity = CL
j. Paragraphing = PARA
k. Introduction = INTRO
l. Conclusion = CONCL
m. Theme = THEME
n. Micro-theme/ Topic Sentence = MIC TH
o. Macro-theme/thesis = MAC TH

5. Cohesion codes

a. Cohesion = Cohesion
b. Cohesive devices = CD
c. Reference = Ref
d. Substitution = Subst
e. Sequencers/discourse markers= DM
f. Linking = Link
g. Subordinators = Subord
h. Coordinators = Coord
i. Conjunctions = Conj
j. Lexical Cohesion = Lexico
k. Unsorted = U

APPENDIX 4: PARTICIPANT BIODATA

20-29 years old	3
30-39	19
40-49	8
50-59	15
60+	5

Table 1: Phase 2 examiners' age distribution

	Full-time	**Part-time**
Less than 2 years	2	1
2-4 years	4	3
5-9 years	21	5
10+ years	22	5

Table 2: Phase 2 examiners' teaching experience

ELICOS	40
AMES	2
High School	2
Uni Prep	9
Unspecified	2

Table 3: Phase 2 examiners' teaching sector

Years as IELTS examiner	Less than2 years	n = 21
	2-4 years	13
	5-9 years	9
	10+ years	8

Table 4: Phase 2 examiners' IELTS experience

APPENDIX 5: PHASE 2 FOLLOW-UP QUESTIONNAIRE

PART A

1. *In general, which criterion do you usually find most difficult to mark? List in order of difficulty (1=most difficult to 4= least difficult or just tick the box for 'all the same level of difficulty').*

   ___ *TR* ___ *CC* ___ *LR* ___ *GRA* □ *all the same level of difficulty*

2. *Which of the band descriptors is the clearest to understand?*

   (1 = clearest to 4 = least clear or just tick the box for 'all equally clear').

   ___ *TR* ___ *CC* ___ *LR* ___ *GRA* □ *all equally clear*

3. *In general, which of the following features of CC play the most significant role in your scoring of this criterion? List in order of significance (1 = most significant to 8 = least significant)*

   ___ Reference

   ___ Substitution

   ___ Paragraphing

   ___ Message/ideas

   ___ Linking words or phrases

   ___ Flow/fluency

   ___ Overall structure

   ___ Logical progression/sequencing of ideas

   ___ Other. Please state: _____

4. *In general, which of the following linguistic features of CC do you usually refer to in your decision making for this criterion? Tick the appropriate box*

	never	seldom	sometimes	very often	always
Reference	☐	☐	☐	☐	☐
Substitution	☐	☐	☐	☐	☐
Paragraphing	☐	☐	☐	☐	☐
Message/ideas	☐	☐	☐	☐	☐
Linking words or phrases	☐	☐	☐	☐	☐
Flow/fluency	☐	☐	☐	☐	☐
Overall structure	☐	☐	☐	☐	☐
Logical progression/ sequencing	☐	☐	☐	☐	☐

Other. Please state: _____

5. *In general, how confident do you feel about rating each of the criteria? Tick one box in each row.*

	not at all confident	not very confident	neither confident nor unconfident	relatively confident	very confident
TR	☐	☐	☐	☐	☐
CC	☐	☐	☐	☐	☐
LR	☐	☐	☐	☐	☐
GRA	☐	☐	☐	☐	☐

6. *How much do you think your rating of coherence and cohesion is affected by the following? Tick one box in each row.*

	not at all	to some extent	a great deal	not applicable
Your experience in teaching writing	☐	☐	☐	☐
Your experience as an IELTS examiner	☐	☐	☐	☐
Your background in Systemic Functional Linguistics	☐	☐	☐	☐
Your IELTS training or standardization	☐	☐	☐	☐
The clarity of the band descriptors	☐	☐	☐	☐
Discussing band descriptors with other examiners	☐	☐	☐	☐

7. *How do you define coherence? Please give a short explanation.*

8. *How do you define cohesion? Please give a short explanation.*

9. *What do 'cohesive devices' refer to? Please give short definition or a list of cohesive devices.*

10. *What does 'substitution' refer to? Please give a short definition.*

11. *What does 'reference' refer to? Please give a short definition.*

12. *Which of the following statements about coherence and cohesion **most closely** represents your point of view? Please tick **one** box:*

☐ A good answer will have a good overall structure with a clear introduction, body and conclusion. The flow of ideas from one sentence to another is not as important as the overall structure.

☐ A good answer flows smoothly and is easy to read. The overall structure is not as important as the flow of ideas from one sentence to another.

---

13. *How useful are the bolded ceilings on paragraphing in your decision-making for marking Task 2. Tick the appropriate box.*

Not at all useful	**not very useful**	**quite useful**	**very useful**	**very useful indeed**
☐	☐	☐	☐	☐

PART B: EXAMINER TRAINING / STANDARDISATION FOR MARKING COHERENCE AND COHESION IN IELTS WRITING TASK 2

1. *How would you rate the examiner training or last standardisation in the assessment of coherence and cohesion?*

poor	**below average**	**average**	**very good**	**excellent**
☐	☐	☐	☐	☐

---

2. *How much do you remember about the training/last standardisation in the assessment of coherence and cohesion?*

nothing	**not much**	**some**	**a great deal**	**everything**
☐	☐	☐	☐	☐

---

3. *Are you aware that at each testing centre, examiners can access standardised scripts for regular revision?* ☐ Yes ☐ No

---

4. *Before you mark scripts, how often do you read through the writing examiners' booklet containing the definitions of each of the criteria?*

☐every time I mark scripts ☐often ☐sometimes ☐rarely ☐never

A number of suggestions have been made for improving examiner training for marking writing, with particular reference to the assessment of coherence and cohesion. Please tick one box to indicate the extent to which you agree with each suggestion.

5. *A detailed analysis of CC in one or two full length scripts for each band level showing all cohesive devices and illustrating for example, their misuse, overuse or omission.*

strongly agree	**agree**	**neither agree nor disagree**	**disagree**	**strongly disagree**
☐	☐	☐	☐	☐

6. *Revision of the CC band descriptors to ensure greater consistency in terminology.*

strongly agree	agree	neither agree nor disagree	disagree	strongly disagree
☐	☐	☐	☐	☐

7. *A glossary of key terms for describing coherence and cohesion with definitions and examples to be included in the instructions for writing examiners.*

strongly agree	agree	neither agree nor disagree	disagree	strongly disagree
☐	☐	☐	☐	☐

8. *More mentoring and feedback in the first year of examining.*

strongly agree	agree	neither agree nor disagree	disagree	strongly disagree
☐	☐	☐	☐	☐

9. *An online question and answer service available for examiners.*

strongly agree	agree	neither agree nor disagree	disagree	strongly disagree
☐	☐	☐	☐	☐

10. *Use of colours to pick out the key features across the bands in the rating scale.*

strongly agree	agree	neither agree nor disagree	disagree	strongly disagree
☐	☐	☐	☐	☐

11. *A list of dos and don'ts for marking scripts (eg don't compare assessments of one script against another) to be included in the examiners' instructions booklet.*

strongly agree	agree	neither agree nor disagree	disagree	strongly disagree
☐	☐	☐	☐	☐

12. *A step-by-step guide to the recommended process (or processes) to follow (eg refresh your memory of the band descriptors before marking) to be included in the instructions for writing examiners booklet.*

strongly agree	agree	neither agree nor disagree	disagree	strongly disagree
☐	☐	☐	☐	☐

13. *Online training materials with exercises for revision and reflection.*

strongly agree	agree	neither agree nor disagree	disagree	strongly disagree
☐	☐	☐	☐	☐

Which **3** improvements do you consider the most important? Add **1 to 3** beside the chosen items.

___ A detailed analysis of CC in one or two full length scripts showing all cohesive ties

___ Revision of the CC band descriptors to ensure greater consistency in terminology

___ A glossary of key terms for describing coherence and cohesion

___ More mentoring and feedback in the first year of examining

___ An online question and answer service available for examiners

___ Use of colours to pick out the key features across the bands

___ A list of dos and don'ts when marking scripts

___ A step-by-step guide to the recommended process to follow.

14. Have you any other comments or suggestions to make for improving training in the marking of CC? Please list below

PART C: BACKGROUND INFORMATION

Please complete the following. Tick the chosen boxes.

Gender: ☐Male ☐Female **Age:** ☐20s ☐30s ☐40s ☐50s ☐60+

1. How many years have you been (ESL) teaching? Tick the box/es

Full-time	☐Less than 2 years	☐2-4 years	☐5-9 years	☐10+ years
Part-time:	☐Less than 2 years	☐2-4 years	☐5-9 years	☐10+ years

*2. In which ESL/TESOL sector have you **mainly** worked? Tick **one** box.*

☐ELICOS ☐AMES ☐Senior high school ☐Other (Please state which sector) _____

*3. At which level do you have the **most** TESOL experience? Tick **one** box.*

☐Elementary ☐Pre-intermediate ☐Intermediate ☐Upper Intermediate ☐Advanced

4. What are your ESL/TESOL qualifications? Tick the chosen boxes:

	Bachelors	Grad Cert	Grad Dip	Masters	PhD
a) ESL/TESOL	☐	☐	☐	☐	☐
b) Applied Linguistics	☐	☐	☐	☐	☐

c) Other. Please state: _____

5. *As part of your qualifications, did you undertake courses that addressed the following? Tick the chosen boxes.*

a)	Discourse analysis	☐ Yes	☐ No
b)	Systemic Functional Linguistics (SFL) text analysis	☐ Yes	☐ No
c)	Formal grammar	☐ Yes	☐ No
d)	How to teach academic writing	☐ Yes	☐ No

6. *Have you ever taught academic writing?* ☐ Yes ☐ No

7. *If yes, how often have you taught an academic writing course? Tick the chosen box.*

☐ 10+ times ☐ 6-9 times ☐4-5 times ☐1-3 times ☐ never

8. *Have you ever taught a dedicated IELTS preparation course?* ☐Yes ☐No

9. If yes, how often have you taught IELTS preparation courses?

☐once ☐2-3 times ☐ 4-5 times ☐More than 5 times

10. *How many years have you been an IELTS writing examiner?*

☐Less than 2 years ☐ 2-4 years ☐ 5-9 years ☐10+ years

11. *On average, how often do you mark IELTS writing scripts?*

☐ Almost every week ☐ twice a month ☐ once a month ☐ once every 2 months ☐ less often than every 2 months

Thank you very much for your help. Please return this questionnaire to Fiona Cotton or Kate Wilson

APPENDIX 6: CORRELATIONS OF SCORES ON CRITERIA WITH STANDARDISED SCORES

Examiner	Overall	TR	CC	LR	GRA
AB1	.873	.826	.873	.929	.891
AB2	.861	.829	.756	.956	.964
AB3	.759	.645	.826	.830	.842
AB4	.849	.789	.955	.910	.906
AB5	.827	.837	.919	.851	.877
AB6	.835	.775	.792	.858	.934
AB7	.866	.905	.866	.890	.872
AB8	.889	.851	.894	.947	.930
AB9	.915	.917	.946	.918	.904
AB10	.842	.841	.856	.870	.837
AB11	.921	.940	.961	.899	.947
AB12	.899	.885	.907	.943	.909
AB13	.862	.870	.752	.915	.899
ABR1	.894	.959	.850	.834	.934
ABR2	.880	.878	.863	.909	.929
ABR3	.801	.857	**.647**	.838	.835
ABR4	.859	.855	.897	.934	.837
ABR5	.879	.910	.835	.932	.923
ABR6	.897	.898	.861	.918	.936
ABR7	.849	.790	.898	.898	.948
ABR8	.859	.923	.882	.812	.886
ABR9	.843	.866	.845	.786	.947
ABR10	.917	.914	.949	.898	.944
ABR11	.922	.918	.891	.964	.906
ABR12	.835	.839	.791	.901	.893
ABR13	.858	.888	.799	.853	.945
ABR14	.896	.883	.875	.917	.947
BA1	.854	.853	.926	.935	.848
BA2	.872	.871	.866	.930	.890
BA3	.837	.787	.758	.876	.976
BA4	.853	.810	.830	.945	.939
BA5	.797	.684	.825	.952	.842
BA6	.871	.935	.935	.823	.867
BA7	.856	.813	.885	.892	.890
BA8	.837	.893	.761	.802	.868
BA9	.864	.832	.904	.889	.902
BA10	.929	.940	.898	.963	.965
BA11	.801	.769	.862	.813	.852
BA12	.893	.883	.909	.908	.880
BA13	.839	.946	.737	.873	.844
BA14	.848	.831	.861	.960	.860
BAR1	.856	.871	.748	.899	.981
BAR2	.826	.871	.762	.847	.865
BAR3	.935	.950	.888	.967	.937
BAR4	.888	.917	.810	.945	.933
BAR5	.928	.948	.917	.953	.928
BAR6	.814	.771	.831	.868	.867
BAR7	.932	.957	.916	.969	.967
BAR8	.881	.824	.883	.953	.919
BAR9	.849	.835	.892	.842	.875
BAR10	.851	.887	.874	.854	.795
BAR11	.917	.958	.918	.957	.893
BAR12	.813	.800	.861	.898	.886
BAR13	**.553**	**.639**	**.492**	**.563**	**.513**
BAR14	.880	.922	.877	.914	.929

APPENDIX 7: CORRELATIONS OF CRITERIA WITH EXAMINER VARIABLES

	TR	CC	LR	GRA	Overall Spearman
Age	-.140	-.130	-.073	.176	-.029
Full-time teaching experience	.247	.060	-.051	-.113	.100
Part-time teaching experience	.505	.279	.068	.094	.411
Level at which most teaching experience	-.099	.010	-.100	.023	-.100
ESL qualifications	.062	-.241	-.002	-.207	-.024
Applied linguistics qualifications	.232	.058	-.087	-.112	.090
Taught academic writing	-.010	-.055	.171	.103	.000
Frequency of teaching academic writing	-.059	-.024	-.263	-.173	-.193
Yrs of experience as IELTS examiner	.057	-.071	-.073	-.069	-.028
IELTS marking experience	.230	.155	-.014	-.135	.122

APPENDIX 8: POINT BISERIAL CORRELATIONS OF DICHOTOMOUS FACTORS WITH CRITERIA

	CC	TR	LR	GRA
Taught academic writing	0.035	-0.012	0.136	0.074
Gender	-0.103	0.031	-0.148	-0.015

APPENDIX 9: EFFECT OF SCRIPTS ON THE RELIABILITY OF EXAMINERS' SCORES

		Sum of Squares	df	Mean Square	F	Sig.
Task Response	Between Groups	.287	3	.096	1.201	.319
	Within Groups	4.066	51	.080		
	Total	4.353	54			
Coherence and Cohesion	Between Groups	.146	3	.049	.654	.584
	Within Groups	3.810	51	.075		
	Total	3.956	54			
Lexical Resource	Between Groups	.080	3	.027	.293	.830
	Within Groups	4.652	51	.091		
	Total	4.733	54			
Grammatical Range and Accuracy	Between Groups	.127	3	.042	.464	.708
	Within Groups	4.633	51	.091		
	Total	4.760	54			
Overall	Between Groups	.034	3	.011	.309	.819
	Within Groups	1.863	51	.037		
	Total	1.897	54			

APPENDIX 10: INDEPENDENT SAMPLES TEST

T tests for overall harshness or leniency against standard scores

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig	t	df	Sig (2-tailed)	Mean Difference	Std Error Difference	Lower	Upper
AB1	Equal variances assumed (EVA)	.313	.577	-.189	94	.851	-.063	.331	-.721	.596
	Equal variances not assumed			-.189	93.613	.851	-.063	.331	-.721	.596
AB2	EVA	2.233	.138	.764	94	.447	.250	.327	-.400	.900
	Not assumed			.764	93.228	.447	.250	.327	-.400	.900
AB3	EVA	3.092	.082	-.329	94	.743	-.104	.316	-.732	.524
	Not assumed			-.329	91.396	.743	-.104	.316	-.732	.524
AB4	EVA	.572	.451	-.229	94	.819	-.083	.363	-.805	.638
	Not assumed			-.229	92.791	.819	-.083	.363	-.805	.638
AB5	EVA	1.913	.170	.783	94	.435	.250	.319	-.384	.884
	Not assumed			.783	91.987	.435	.250	.319	-.384	.884
AB6	EVA	.000	.996	.183	94	.855	.063	.342	-.616	.741
	Not assumed			.183	94.000	.855	.063	.342	-.616	.741
AB7	EVA	.000	.993	.300	94	.765	.104	.348	-.586	.795
	Not assumed			.300	93.896	.765	.104	.348	-.586	.795
AB8	EVA	.019	.889	-.978	94	.330	-.333	.341	-1.010	.343
	Not assumed			-.978	93.995	.330	-.333	.341	-1.010	.343
AB9	EVA	.604	.439	-1.233	94	.221	-.458	.372	-1.196	.280
	Not assumed			-1.233	91.843	.221	-.458	.372	-1.196	.280
AB10	EVA	.823	.367	.408	94	.684	.146	.357	-.564	.856
	Not assumed			.408	93.331	.684	.146	.357	-.564	.856
AB11	EVA	4.456	.037	-1.229	94	.222	-.479	.390	-1.253	.295
	Not assumed			-1.229	89.269	.222	-.479	.390	-1.254	.295
AB12	EVA	1.348	.249	.445	94	.657	.167	.374	-.576	.910
	Not assumed			.445	91.519	.657	.167	.374	-.577	.910
AB13	EVA	4.653	.034	1.208	94	.230	.375	.311	-.242	.992
	Not assumed			1.208	89.936	.230	.375	.311	-.242	.992
ABR1	EVA	1.402	.239	1.160	94	.249	.375	.323	-.267	1.017
	Not assumed			1.160	92.667	.249	.375	.323	-.267	1.017
ABR2	EVA	.012	.915	.582	94	.562	.208	.358	-.502	.919
	Not assumed			.582	93.297	.562	.208	.358	-.502	.919
ABR3	EVA	5.130	.026	.137	94	.892	.042	.305	-.564	.647
	Not assumed			.137	88.097	.892	.042	.305	-.564	.647
ABR4	EVA	.889	.348	.260	94	.795	.083	.320	-.552	.719
	Not assumed			.260	92.160	.795	.083	.320	-.552	.719
ABR5	EVA	.175	.676	.377	94	.707	.125	.332	-.534	.784
	Not assumed			.377	93.648	.707	.125	.332	-.534	.784
ABR6	EVA	6.105	.015	1.206	94	.231	.458	.380	-.296	1.213
	Not assumed			1.206	90.709	.231	.458	.380	-.297	1.213
ABR7	EVA	.008	.931	.122	94	.903	.042	.341	-.634	.718
	Not assumed			.122	93.993	.903	.042	.341	-.634	.718
ABR8	EVA	2.262	.136	.842	94	.402	.271	.322	-.368	.909
	Not assumed			.842	92.409	.402	.271	.322	-.368	.909
ABR9	EVA	3.627	.060	-.424	94	.673	-.167	.393	-.948	.614
	Not assumed			-.424	88.702	.673	-.167	.393	-.948	.615
ABR10	EVA	.406	.525	.340	94	.734	.125	.367	-.605	.855
	Not assumed			.340	92.347	.734	.125	.367	-.605	.855
ABR11	EVA	.314	.576	.821	94	.413	.271	.330	-.384	.925
	Not assumed			.821	93.462	.413	.271	.330	-.384	.926

ABR12	Equal variances assumed (EVA)	3.182	.078	2.081	94	.040	.646	.310	.030	1.262
	Equal variances not assumed			2.081	89.864	.040	.646	.310	.029	1.262
ABR13	EVA	.270	.604	-1.224	94	.224	-.438	.358	-1.147	.272
	Not assumed			-1.224	93.323	.224	-.438	.358	-1.147	.272
ABR14	EVA	.526	.470	.630	94	.530	.229	.364	-.493	.951
	Not assumed			.630	92.756	.530	.229	.364	-.493	.951
BA1	EVA	2.585	.111	.264	94	.793	.083	.316	-.544	.711
	Not assumed			.264	91.307	.793	.083	.316	-.544	.711
BA2	EVA	1.716	.193	1.901	94	.060	.604	.318	-.027	1.235
	Not assumed			1.901	91.719	.060	.604	.318	-.027	1.235
BA3	EVA	.368	.546	.507	94	.613	.167	.329	-.486	.819
	Not assumed			.507	93.372	.613	.167	.329	-.486	.819
BA4	EVA	.000	.992	-.122	94	.903	-.042	.341	-.718	.635
	Not assumed			-.122	93.995	.903	-.042	.341	-.718	.635
BA5	EVA	5.648	.020	-.069	94	.945	-.021	.301	-.618	.576
	Not assumed			-.069	86.523	.945	-.021	.301	-.618	.577
BA6	EVA	.566	.454	-.062	94	.950	-.021	.334	-.684	.643
	Not assumed			-.062	93.789	.950	-.021	.334	-.684	.643
BA7	EVA	5.928	.017	.836	94	.405	.250	.299	-.344	.844
	Not assumed			.836	85.895	.406	.250	.299	-.345	.845
BA8	EVA	.457	.501	1.233	94	.221	.417	.338	-.255	1.088
	Not assumed			1.233	93.949	.221	.417	.338	-.255	1.088
BA9	EVA	2.005	.160	-.977	94	.331	-.375	.384	-1.137	.387
	Not assumed			-.977	90.148	.331	-.375	.384	-1.138	.388
BA10	EVA	1.945	.166	-.442	94	.660	-.167	.377	-.916	.582
	Not assumed			-.442	91.097	.660	-.167	.377	-.916	.583
BA11	EVA	.445	.507	.312	94	.755	.104	.333	-.558	.766
	Not assumed			.312	93.751	.755	.104	.333	-.558	.766
BA12	EVA	.000	.984	-.248	94	.804	-.083	.335	-.749	.583
	Not assumed			-.248	93.854	.804	-.083	.335	-.749	.583
BA13	EVA	2.225	.139	-1.994	94	.049	-.625	.313	-1.247	-.003
	Not assumed			-1.994	90.702	.049	-.625	.313	-1.248	-.002
BA14	EVA	.576	.450	.503	94	.616	.167	.331	-.491	.825
	Not assumed			.503	93.610	.616	.167	.331	-.491	.825
BAR1	EVA	.894	.347	-1.071	94	.287	-.396	.369	-1.129	.338
	Not assumed			-1.071	92.114	.287	-.396	.369	-1.130	.338
BAR2	EVA	.990	.322	.950	94	.344	.313	.329	-.340	.965
	Not assumed			.950	93.376	.344	.313	.329	-.340	.965
BAR3	EVA	.480	.490	-.314	94	.754	-.104	.331	-.762	.554
	Not assumed			-.314	93.604	.754	-.104	.331	-.762	.554
BAR4	EVA	.138	.711	2.436	94	.017	.813	.333	.150	1.475
	Not assumed			2.436	93.751	.017	.813	.333	.150	1.475
BAR5	EVA	.053	.818	.839	94	.404	.292	.348	-.399	.982
	Not assumed			.839	93.903	.404	.292	.348	-.399	.982
BAR6	EVA	1.760	.188	.197	94	.844	.063	.317	-.568	.693
	Not assumed			.197	91.647	.844	.063	.317	-.568	.693
BAR7	EVA	.487	.487	-1.321	94	.190	-.479	.363	-1.199	.241
	Not assumed			-1.321	92.859	.190	-.479	.363	-1.199	.241
BAR8	EVA	.533	.467	.190	94	.850	.063	.328	-.590	.715
	Not assumed			.190	93.342	.850	.063	.328	-.590	.715
BAR9	EVA	2.113	.149	-.397	94	.692	-.125	.315	-.750	.500
	Not assumed			-.397	91.049	.692	-.125	.315	-.750	.500
BAR10	EVA	.216	.643	-.428	94	.669	-.146	.340	-.822	.530
	Not assumed			-.428	93.992	.669	-.146	.340	-.822	.530
BAR11	EVA	3.911	.051	-.613	94	.541	-.188	.306	-.795	.420
	Not assumed			-.613	88.428	.541	-.188	.306	-.795	.420
BAR12	EVA	.028	.867	-1.033	94	.304	-.354	.343	-1.035	.327
	Not assumed			-1.033	93.997	.304	-.354	.343	-1.035	.327
BAR13	EVA	4.395	.039	-1.747	94	.084	-.729	.417	-1.558	.100
	Not assumed			-1.747	84.816	.084	-.729	.417	-1.559	.101
BAR14	EVA	3.001	.086	-.199	94	.842	-.063	.314	-.685	.560
	Not assumed			-.199	90.758	.842	-.063	.314	-.685	.560

T tests of CC against standard scores for harshness or leniency

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig	t	df	Sig (2-tailed)	Mean Difference	Std Error Difference	Lower	Upper
AB1	Equal variances assumed (EVA)	.636	.428	-.079	58	.937	-.042	.527	-1.098	1.014
	Equal variances not assumed			-.086	19.118	.932	-.042	.482	-1.051	.968
AB2	EVA	4.776	.033	.735	58	.465	.375	.510	-.646	1.396
	Not assumed			.948	25.909	.352	.375	.395	-.438	1.188
AB3	EVA	5.840	.019	.247	58	.806	.125	.506	-.888	1.138
	Not assumed			.335	29.112	.740	.125	.373	-.638	.888
AB4	EVA	.000	.986	.230	58	.819	.125	.544	-.963	1.213
	Not assumed			.226	16.595	.824	.125	.553	-1.044	1.294
AB5	EVA	2.179	.145	.716	58	.477	.375	.524	-.674	1.424
	Not assumed			.805	19.979	.430	.375	.466	-.597	1.347
AB6	EVA	.647	.425	.396	58	.694	.208	.526	-.845	1.262
	Not assumed			.437	19.409	.667	.208	.476	-.788	1.204
AB7	EVA	.014	.906	.677	58	.501	.375	.554	-.734	1.484
	Not assumed			.631	15.652	.537	.375	.594	-.887	1.637
AB8	EVA	.045	.833	-.388	58	.699	-.208	.537	-1.283	.867
	Not assumed			-.397	17.425	.696	-.208	.525	-1.314	.897
AB9	EVA	1.355	.249	-.791	58	.432	-.458	.579	-1.618	.701
	Not assumed			-.665	14.244	.516	-.458	.689	-1.933	1.017
AB10	EVA	.006	.938	.384	58	.703	.208	.543	-.879	1.295
	Not assumed			.379	16.672	.710	.208	.550	-.954	1.371
AB11	EVA	1.458	.232	-1.091	58	.280	-.625	.573	-1.772	.522
	Not assumed			-.938	14.510	.364	-.625	.666	-2.050	.800
AB12	EVA	.105	.747	.077	58	.939	.042	.544	-1.047	1.131
	Not assumed			.075	16.551	.941	.042	.555	-1.131	1.215
AB13	EVA	4.590	.036	.735	58	.465	.375	.510	-.646	1.396
	Not assumed			.948	25.909	.352	.375	.395	-.438	1.188
ABR1	EVA	.845	.362	.393	58	.695	.208	.530	-.852	1.268
	Not assumed			.423	18.674	.677	.208	.492	-.823	1.240
ABR2	EVA	.043	.836	.850	58	.399	.458	.539	-.621	1.538
	Not assumed			.857	17.123	.403	.458	.535	-.669	1.586
ABR3	EVA	4.075	.048	.245	58	.807	.125	.510	-.895	1.145
	Not assumed			.318	26.212	.753	.125	.393	-.682	.932
ABR4	EVA	.744	.392	.080	58	.937	.042	.524	-1.007	1.090
	Not assumed			.089	19.979	.930	.042	.466	-.930	1.013
ABR5	EVA	.514	.476	.555	58	.581	.292	.525	-.760	1.343
	Not assumed			.618	19.650	.544	.292	.472	-.694	1.277
ABR6	EVA	3.345	.073	.654	58	.515	.375	.573	-.772	1.522
	Not assumed			.563	14.510	.582	.375	.666	-1.050	1.800
ABR7	EVA	1.200	.278	-.558	58	.579	-.292	.523	-1.338	.755
	Not assumed			-.634	20.294	.533	-.292	.460	-1.251	.668
ABR8	EVA	.915	.343	.707	58	.483	.375	.531	-.687	1.437
	Not assumed			.754	18.456	.460	.375	.497	-.668	1.418
ABR9	EVA	1.943	.169	-.072	58	.943	-.042	.582	-1.208	1.124
	Not assumed			-.060	14.120	.953	-.042	.700	-1.542	1.459
ABR10	EVA	.114	.737	-.074	58	.941	-.042	.560	-1.164	1.080
	Not assumed			-.067	15.189	.947	-.042	.620	-1.361	1.278
ABR11	EVA	.363	.549	.236	58	.814	.125	.530	-.937	1.187
	Not assumed			.252	18.536	.804	.125	.495	-.913	1.163
ABR12	EVA	.670	.416	1.512	58	.136	.792	.523	-.256	1.840
	Not assumed			1.707	20.095	.103	.792	.464	-.175	1.759

ABR13	Equal variance assumed (EVA)	.045	.833	-.528	58	.600	-.292	.553	-1.398	.815
	Equal variance not assumed			-.494	15.737	.628	-.292	.590	-1.544	.961
ABR14	EVA	.131	.719	.835	58	.407	.458	.549	-.641	1.558
	Not assumed			.796	16.050	.437	.458	.576	-.761	1.678
BA1	EVA	3.171	.080	.567	58	.573	.292	.515	-.739	1.322
	Not assumed			.693	23.191	.495	.292	.421	-.579	1.162
BA2	EVA	1.235	.271	1.841	58	.071	.958	.521	-.084	2.000
	Not assumed			2.130	20.944	.045	.958	.450	.023	1.894
BA3	EVA	.204	.654	.702	58	.485	.375	.534	-.694	1.444
	Not assumed			.732	17.869	.474	.375	.512	-.702	1.452
BA4	EVA	.237	.628	.078	58	.938	.042	.537	-1.034	1.117
	Not assumed			.079	17.366	.938	.042	.527	-1.068	1.151
BA5	EVA	2.915	.093	-.081	58	.936	-.042	.514	-1.070	.987
	Not assumed			-.100	23.761	.921	-.042	.415	-.898	.815
BA6	EVA	.115	.736	-.537	58	.593	-.292	.543	-1.379	.795
	Not assumed			-.530	16.672	.603	-.292	.550	-1.454	.871
BA7	EVA	1.139	.290	.720	58	.474	.375	.520	-.667	1.417
	Not assumed			.835	20.990	.413	.375	.449	-.559	1.309
BA8	EVA	1.855	.179	.877	58	.384	.458	.522	-.587	1.504
	Not assumed			1.000	20.417	.329	.458	.458	-.496	1.413
BA9	EVA	.484	.489	-.368	58	.714	-.208	.566	-1.342	.925
	Not assumed			-.325	14.845	.750	-.208	.642	-1.578	1.161
BA10	EVA	2.108	.152	-.362	58	.719	-.208	.576	-1.361	.944
	Not assumed			-.308	14.388	.762	-.208	.676	-1.655	1.239
BA11	EVA	1.508	.224	.399	58	.692	.208	.523	-.838	1.255
	Not assumed			.453	20.294	.656	.208	.460	-.751	1.168
BA12	EVA	.007	.936	.078	58	.938	.042	.537	-1.034	1.117
	Not assumed			.079	17.366	.938	.042	.527	-1.068	1.151
BA13	EVA	3.111	.083	-1.556	58	.125	-.792	.509	-1.810	.227
	Not assumed			-2.035	26.744	.052	-.792	.389	-1.590	.007
BA14	EVA	.619	.435	.393	58	.695	.208	.530	-.852	1.268
	Not assumed			.423	18.674	.677	.208	.492	-.823	1.240
BAR1	EVA	.006	.939	-.534	58	.595	-.292	.546	-1.385	.802
	Not assumed			-.517	16.323	.612	-.292	.564	-1.485	.902
BAR2	EVA	2.179	.145	1.034	58	.305	.542	.524	-.507	1.590
	Not assumed			1.163	19.979	.259	.542	.466	-.430	1.513
BAR3	EVA	1.982	.165	-.081	58	.936	-.042	.517	-1.077	.994
	Not assumed			-.096	22.186	.924	-.042	.433	-.939	.855
BAR4	EVA	.363	.549	1.493	58	.141	.792	.530	-.270	1.853
	Not assumed			1.598	18.536	.127	.792	.495	-.247	1.830
BAR5	EVA	.012	.913	.535	58	.595	.292	.545	-.800	1.383
	Not assumed			.521	16.420	.609	.292	.560	-.893	1.476
BAR6	EVA	1.927	.170	.401	58	.690	.208	.519	-.831	1.248
	Not assumed			.470	21.377	.643	.208	.444	-.713	1.130
BAR7	EVA	.015	.903	-1.142	58	.258	-.625	.547	-1.721	.471
	Not assumed			-1.100	16.215	.288	-.625	.568	-1.828	.578
BAR8	EVA	1.569	.215	.555	58	.581	.292	.525	-.760	1.343
	Not assumed			.618	19.650	.544	.292	.472	-.694	1.277
BAR9	EVA	1.235	.271	-.080	58	.936	-.042	.521	-1.084	1.000
	Not assumed			-.093	20.944	.927	-.042	.450	-.977	.894
BAR10	EVA	.331	.567	-.233	58	.817	-.125	.537	-1.201	.951
	Not assumed			-.237	17.366	.815	-.125	.527	-1.235	.985
BAR11	EVA	4.432	.040	-.247	58	.806	-.125	.506	-1.139	.889
	Not assumed			-.333	28.696	.742	-.125	.376	-.894	.644
BAR12	EVA	.015	.903	-1.142	58	.258	-.625	.547	-1.721	.471
	Not assumed			-1.100	16.215	.288	-.625	.568	-1.828	.578
BAR13	EVA	.899	.347	-1.195	58	.237	-.708	.593	-1.895	.478
	Not assumed			-.963	13.781	.352	-.708	.735	-2.288	.871
BAR14	EVA	.636	.428	-.079	58	.937	-.042	.527	-1.098	1.014
	Not assumed			-.086	19.118	.932	-.042	.482	-1.051	.968

APPENDIX 11: EXAMINERS' SUGGESTIONS AND COMMENTS ABOUT TRAINING IN CC

- More comprehensive initial training; online support
- More practice marking scripts; trainer to use OHPs to highlight/demonstrate why the script would receive a certain band score
- Review of Bands 8 and 9 in CC to create a clearer distinction between the two bands
- I don't think it's that bad! I think people complain because they don't know what cohesive devices are and have a lack of knowledge themselves about cohesion and grammar. I can work with what is here already, however, some of the examiners may not be skilled enough to manage their jobs adequately. While there is some confusion in the bands, teachers must train themselves and be aware of grammar and English language rules, in order to test and mark effectively.
- Applying bands to analysed h/work texts given during the training
- A bit more detailed, and more aspects taken into consideration will help us be more accurate in marking. This is something that I suggest it (sic) happens in all 4 band descriptors because some of them are very broad.
- More practice at marking papers to identify all aspects of CC
- More collaborative modes at training/standardisation sessions. Time to reflect with other examiners (depends on one's learning style).
- The writing model which is being used by IELTS assumes that each category (TR, CC, LR, GRA) are of equal value or weighting.
- Need more workshops (half yearly). More opportunity to mark writing – always on speaking! Opportunity to cross check markings with other examiners ie 1) marking in pairs – time consuming but more adequate. 2) senior examiner be present during marking
- A couple of sample scripts with comprehensive highlighting of cohesive features would be very helpful.
- Train examiners in each component of CC separately. Coherence first and then cohesion.
- Make band descriptors clearer and more specific
- Have more time to spend discussing the band descriptors in update training sessions. Clarify range of cohesive devices
- Example of flat 6 writing for each task available in examiners' booklet
- "Substitution" and "reference" need clarification
- I think the above three should be implemented [analysis of CC in several scripts, glossary of key terms and an online question and answer service] and then it would be much clearer.
- Allow trainee examiners to take home scripts to rate during training for more specific, individualised feedback
- I believe any online or face to face training needs to be re-numerated as everyone is time poor and the pay for marking/examining on the day is quite limited in that you don't get paid (or very little) for preparation. There is, in other words, very little time to read the examiner instruction booklet.
- Examiners to be given examples of scripts with processes used by examiners to come to these results.
- People forget re articles and 'double dip' with the GRA section at times. I have to keep checking this myself.
- A clearer distinction needs to be made between coherence and cohesion, as well as guidelines for their assessment. At times, I find the distinction between TR and CC quite blurred – greater focus on what is specifically required for each would be beneficial.