Investigating Raters' Use of Analytic Descriptors in Assessing Writing

Wu Siew Mei
National University of Singapore

ABSTRACT

The rating process of written tests has been fraught with various areas of difficulty in spite of a body of research examining these issues in the last decade. This study explores the rating processes of four markers using an analytic rating scale to mark a set of 16 undergraduates' scripts at the National University of Singapore. The three main questions are:

(1) What kind/s of sequence/s do raters go through as they rate a set of scripts based on an analytic rating scale?
(2) What are some factors contributing to rater indecision when using the descriptors?
(3) How do raters make decisions regarding the award of marks when there is a perceived difficulty in the application of descriptor guidelines?

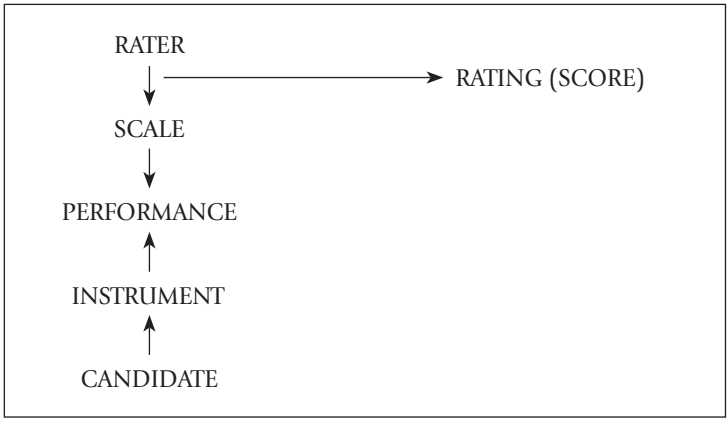
The study mainly uses think-aloud protocols that the raters are trained to produce. Raters' comments along the margin were also inspected and 20-minute interviews were conducted with each rater to provide detailed insights into the rating process. The study shows that though raters had some common rater behavior in their marking sequence, there were also major differences in the way they dealt with difficult areas especially in assessing the content of essays. Strategies used to manage junctures of indecision seem to vary too.

KEYWORDS: *Writing assessment; Analytic descriptors; Inter-rater reliability*

The direct test of writing "where test takers actually produce a sample of writing" (Weigle, 2002, p. 58) has remained a standard component in many international English language assessment instruments in spite of concerns about its value in accurately reflecting test takers' real writing proficiency levels. Milanovic, Saville, and Shen (1996) reiterate "the faith in their validity and good impact on their teaching" (p.92) that has seen direct written tests used as central components in international examinations at both L1 and L2 levels, including the Cambridge examinations. Perhaps, concerns over the extent to which ratings reflect the raters' subjective interpretations of test takers' writing ability rather than their actual writing competency (Lumley, 2000) are best summed up by Cumming, Kanto, and Powers (2001, p. 3) who aptly point out the abject contrast between the "simplicity of the holistic scoring method, and the rating scales that typically accompany it" and the "complex, richly informed judgments of skilled

human raters to interpret the value of the worth of students' writing abilities".

The complexity of rating direct written tests can be attributed to the interaction of different variables including the background characteristics of the test taker, the language/or behaviour of the rater and the functioning of assessment criteria, rating scales and scoring rubrics. Such tests are open to various errors not least because of the human element involved in each stage in the rating process of these tests. Hamp-Lyons (1990) points out the detrimental impact of this openness to errors on the reliability of the test for the assessment of writing ability. McNamara (1996) represents the array of factors in the scoring procedure as such:



McNamara (1996)

Calls for investigation into the rating process have been sounded by several researchers including Hamp-Lyons (1991), Weigle (1994, 1998), Cumming (1997) and Kroll (1998).

In the last decade or so, the reliability of writing performance assessment has been reinstated through a variety of approaches including a battery of tests for its "technical soundness" (Milanovic et al., 1996, p.92), the training of raters and better specification of scoring criteria and task (Lumley, 2002). However, Milanovic et al. (1996) highlight the need for more attention to be given to rater behaviour in the marking process as the rater was recognized as "one of the main sources of measuring error in assessing a candidate's performance" in the context of research at Cambridge. They call for a "better understanding of the value, decision-making behaviour and even the idiosyncratic nature of the judgements markers make" (Milanovic et al., 1996, p.92).

This concern parallels Huot's (1990) earlier comments that " . . . little is known about the way raters arrive at these decisions . . . we have little or no information on what role scoring procedures play in the reading and rating process."

Lumley (2002) highlights one specific area of concern in the rating process: the superficiality of rating scales as compared to the complexities that operate in written texts and the subjectivity involved in the interpretation of these scales. Such a mismatch needs to be further examined in a detailed manner so that rating

instruments and processes can be fine tuned to minimize unfair practices in the assessment of writing competencies.

The written test is a major component in many tertiary English language tests as such an open ended response is deemed to be a good source to infer students' language abilities for decision making purposes. The results of such tests can affect the range of academic career choices open to the test takers and the modules they are allowed to offer. In some cases, the stakes are higher in that the candidates' performance has an impact on the choice of institutions they can enroll in. In high-stake tests, the results "have a significant impact on the lives of individuals or on programs, and [the effects] are not easily reversed" (Weigle, 2002, p.41). This is a reason to understand rater behaviour and to minimize the level of subjectivity in the rating process.

In the context of the present study, a better understanding of the strategies used by the raters will provide insights into how the currently used descriptors can be fine tuned to minimize subjectivity in the scoring criteria. In some instances, new criteria may need to be added to cater to lacks in the descriptors. In other cases, certain criteria may have to be removed if they do not prove useful to raters or in a worst case scenario, contribute to confusion in the marking process. At a more general level, such rater strategies can be useful information to developers of other analytic scoring descriptors as this is a rather common procedure in marking written tests.

Additionally, this study will also provide useful information for the training of larger groups of raters as strategies can be identified, described, and critiqued in a more concrete manner. Common patterns that facilitate consistency can be encouraged while the rationale for idiosyncratic behaviour can be closely investigated and appraised for their impact. This will help raters to have a firmer grasp of what should be done especially in problematic junctures in the marking process. Ultimately, such steps can only lead to a higher level of self awareness, accuracy, and professionalism in the rating process, a central aspect of language teaching.

The objectives of this study are:

- To identify the range of rater behaviour that are common in the use of rating scales
- To identify idiosyncratic rater practices
- To understand the motivation of these practices and their possible impact on scoring
- To ascertain the levels of emphasis placed on various features described in the rating scale
- To identify strategies of resolution used by the raters when they meet with a problematic situation in using the scale (e.g., conflict resolution)
- To suggest implications of rater behaviour on the training of raters.

Related Research

Direct tests of writing as described by Hamp-Lyons (1991) consist of the following features:

- A continuous text of at least 100 words
- Writers respond to a set of instructions or prompt but with leeway given for different responses
- Text is read by at least one but normally two or more trained raters
- Judgment is tied to common yardstick
- Judgment is expressed in numbers.

These tests are usually done within a limited time frame and the topic is known only at the point of examination.

Weigle (2002) identifies three main sources of variation in the assessment of such direct writing tests: candidates' ability, choice of task, and raters. She stresses the need for examiners to pay close attention to the latter two as Linacre (1989) points out that it has been recognized for at least a century that rater variability is extensive accounting for "between one and two-thirds of the variability in a set of scores, that is as much as differences in ability among candidates" (Weigle, 2002, p. 122).

The awareness of sources of rater variability in scoring has resulted in these practices to contain the level of inconsistency:

- Carefully worded definitions and descriptions of performance criteria with examples illustrating characteristic performance provided
- The use of experienced raters who are carefully trained
- Double marking procedures for reconciling disagreements.

Also, research has been ongoing to investigate the interaction of various factors in both holistic and analytic assessment of essays. Generally, there are two main foci in such investigations. These include a consideration of

- the attributes of texts raters pay attention to in the evaluation process
- background rater characteristics and their effects on the reading process and ultimately the scoring of texts.

Many of these studies make use of verbal protocol analysis which is not without its fair share of criticisms on its validity (See for example Stratman & Hamp-Lyons, 1994). Think aloud protocols essentially require raters to verbalise their thought processes into a taping device as they mark designated scripts. These verbalizations are subsequently transcribed to allow analysis and interpretation by the researcher. The technique gains its validity mostly on the basis of Ericsson and Simon's (1993) work. They point out that though much of a rater's thought processes may remain embedded, the accessibility to short term memory using such a technique can provide useful insights to some extent.

Earlier research into the validity of the scoring of written assessments mainly involved correlation studies of traits that are characteristic of high and low score essays (see for example Homburg, 1984 & Sparks, 1988). Later research work, in response to the call for investigation into the rating process, has resulted in studies that examine rater behaviour in various assessment contexts. One major area of focus in these studies involves the identification of criteria that raters use to rate the essay. (See for example Cumming, 1989; Vaughan, 1991; Weigle, 1994;

Milanovic et al., 1996; Sakyi, 2001). These investigations mainly use concurrent and retrospective verbal protocol analysis to provide insights into what actually goes on in raters' minds when they make critical decisions and the specific criteria used as they read and scored essays. These studies show that experienced raters are clear in the criteria that they use to assess the essays and in their rating strategies.

Cumming (1989) identifies 28 interpretation strategies that experienced raters use and these strategies can be classified into three major categories that form the basis for the award of scores: substantive content, language use, and rhetorical organization. Vaughan (1991) shows that experienced raters agreed on rating criteria in scoring guides but were likely to use their own style when the essays were incongruent with set standards in the guide. Sakyi (2001) identifies four distinct reading styles in the reading processes of experienced raters. They focused on errors in the text, essay topics, presentation of ideas, personal reaction to text, and scoring guides. The study also found that raters who make a conscious effort to follow the scoring guide were more likely to focus on one or two features to differentiate abilities in writing when they had to award a single score at the end of the marking process.

These studies collectively explore decisions and criteria used in the holistic marking of written tests where marks were awarded based on an overall impression of the quality of writing. However, what is lacking is the need to better understand rater behaviour in the analytic marking process. Scant attention has been given to the raters' application of rating scales in the rating process. In analytic marking, raters mark according to a list of set descriptors that categorically describe the expected quality of different aspects of the essay (e.g., content, organization, cohesion, grammar) for bands of marks to be awarded to different aspects of the writing e.g., Content, Language, and Organisation.

There is a debate as to whether the analytic process necessarily results in more accurate rating. Charney (1984) recommends that reliability is best achieved when rating is quick and impressionistic as in-depth consideration can only lead to further interpretation and thus higher inconsistencies. Huot (1993), however, argues that this is not necessarily so. Weigle (2002, p. 73) provides evidence from various studies including Bauer (1981) to argue for the relatively higher reliability of analytic scoring. The selection of one test scale over another is not always clear. Moreover, Weigle (2002) highlights the fact that "there has been surprisingly little research on the effects of different scale types on outcomes" (p.72). Therefore, the way the raters actually use these descriptors and the kinds of interpretations they make at various junctures of the process are some aspects of rater behaviour that warrant further investigation.

McNamara (1996) outlines variability amongst raters in these ways:

- Two raters are different in their overall leniency
- Raters may be particularly harsh or lenient towards certain groups and not others or towards certain questions and not others
- Raters may have different interpretations of the descriptor that they are using
- Raters may differ in their consistency or inconsistency.

Such sources of variability amongst raters become pertinent especially in the context of a study by Lumley (2002, 2005) which investigates the analytic rating process and the interaction of descriptor and text in a large-scale writing test. Lumley (2002) argues that the main function of descriptors and training is to provide an avenue for raters to channel their diverse reactions to texts into narrower, more manageable statements that are crafted according to institutional demands and expectations. It helps ". . . raters to articulate and justify their rating decisions in terms of what the institution requires, in the interest of reliable, orderly, and categorized ratings" (Lumley, 2002, p. 267).

Lumley's study analyses the sequence of rating, the interpretations raters made of scoring categories and the difficulties that they face in the use of the descriptors. The study shows that raters basically follow a similar sequence of rating but what is not clear is how the scale content affects marks awarded for the quality of the written texts. The investigation highlights the tension between raters' overall impression and the specific wordings in the rating scale and in some instances, the fact that the scale may not address other possible features in the essays that may prove problematic. Raters are then forced to use other strategies to cope with these conflicts. However, despite these areas of obscurity, the study also shows that consistency in rating can be achieved with adequate training and additional guidelines to supplement the list of descriptors used.

The present investigation draws on the approach used in Lumley's (2002) study in terms of aspects of the methodology employed. Lumley's study focuses on similarities and differences in the rating behavior of a small group of highly experienced raters with the overaching aim of establishing generalisable patterns of rater behavior. In this study, the research interest is similar but narrower than Lumley's. Also, it is a smaller scale investigation to delve more in-depth into the nature and rationale for deviant use of descriptors and the manner in which indecisions are resolved when the typical sequence of use of descriptors is broken. The research questions in this study are:

- What are some characteristic practices amongst experienced raters in the application of rating scales to written assessments?
- What are some idiosyncratic rating practices that may contribute to less than consistent use of rating scales amongst experienced raters?
- What are some strategies that experienced raters use in the resolution of areas of obscurities that are not catered for by rating scales?
- What are some factors that contribute to the raters' final decision in the scoring process?

The study uses think aloud protocols and retrospective interviews with a selected group of raters to elicit their thought processes in the rating of essays.

Methodology

Raters and Scripts

Four raters were selected based on their similarity in terms of qualifications and years of experience in teaching. These raters had also marked at least three rounds

of the annual large-scale English placement test. The scripts consisted of essays written by students enrolled in the English for Academic Purposes module in that semester. The task prompt to the essay was as follows:

> The Writing Task
>
> Reflect on the ideas generated from the pre-reading questions and from reading the two passages; then write **an essay of about 500 words** in response to the following question:
>
> *One's true potential and character emerge only in competitive situations. Do you agree with this statement? Support and justify your stand.*
>
> Note that:
> - You should use personal opinions, clear examples, and well-organised points to support your answer.
> - You **may use ideas** from either of the passages but you **must not copy sentences directly** from them.
>
> Your reader is a university lecturer.

The post-course test, from which the essays were drawn, was marked analytically with the guide of a descriptor (see Appendix A for a sample descriptor) and a rater training session prior to the marking process. Essays were assessed for the three major categories of Content, Language, and Organisation with bands ranging from 0 to 6, 0 being the lowest and 6, the highest band. Each band in each category has an accompanying profile description which essentially stipulates key features to focus on with respect to particular categories and their corresponding bands (see Appendix A).

Essentially, the raters went through the following rounds of scoring in the study:

(a) Round 1—raters scored two scripts each to provide a form of simulated marking environment. The intent here was for the researcher/trainer to ascertain that the raters are not those who might be extreme outliers in their level of leniency or strictness in their marking. It also provided an opportunity for the raters to generally review the requirements of the marking process, although they had all experienced similar marking processes in the course of their professional duties.

(b) Round 2—raters were given four scripts to assess and write comments along the margin as in a non-investigative context. In this round of marking, the items listed below were carried out:
   i) Raters reviewed the question prompt and explained their understanding of the requirements of the prompt.
   ii) Raters looked through the rating scale to review certain terms such as "meaning obscured".
   iii) Raters provided indications of their scores in each category and the reasons for the scores were provided. In the cases of large discrepancies, negotiation amongst the raters resulted in a level of consensus.

(c) Round 3—raters marked another set of four essays each (taken from the same

set of post course scripts) providing a think-aloud analysis—instructions for protocol given earlier (see Appendix B for instructions to raters on think-aloud protocols).

(d) An interview was conducted with each of the raters to elicit retrospective comments at junctures where the analyst deemed appropriate.

A range of qualitative techniques such as interviews, retrospective recall or verbal protocol analysis provide empirically-based descriptions and evaluations of written products that help us understand rater behavior in more specific ways. These insights can be fed back directly into pedagogical measures such as rater training or better development of rating scales and rubrics.

The procedure for obtaining the think-aloud comments draws largely from those outlined by Ericsson and Simon (1993). Subsequently, an interview was conducted with each of the raters to elicit retrospective comments at junctures where the analyst deemed appropriate and necessary to obtain further clarification (see Appendix C for a sample of interview questions).

Coding the Data

A broad orthographic transcription was presented for each of the raters with appropriate conventions to indicate, for instance, the source of rater comments. The basic principle guiding the coding of raters' think-aloud data was to best capture the kinds of information heeded by the raters as the verbal protocols were produced. Coding categories were not formed based on fixed assumptions or theoretical understanding of what forms typical rating behaviour should take. On the contrary, the coding categories were developed based on observations of the consistency of the categories with the rating scale and the purposes of the research. It was a precarious balance between maintaining generalisability and specificity so as to avoid making interpretations that are too vague to be useful or capturing merely idiosyncratic behaviour (Green, 1997). Drawing on Lumley (2002), the coding categories for this study were developed along these broad behavioural patterns:

- Management behaviours
- Reading behaviours
- Rating behaviours.

The unit of analysis decided upon was a "single or several utterances with a single aspect of the event as the focus" (Green, 1997, cited in Brown, 2000). Such a unit may consist of one clause or many clauses but each of them centres on a dominant event as focus. Examples A and B provide two such examples:

Example A

So I think that his first paragraph has a focus thesis point. (InF4/9)

Res ⟶ +ve ⟶ organisation
(Res – Response ; +ve – positive; organization – organization category)

Table 1
Coding Schemes Used

Code	Code meanings
Read	Rater reading text
Manage ⟶ Procedure	Rater engaged in management of the marking procedure, e.g., I will mark S100 first.
Res ⟶ –ve ⟶ language Res ⟶ +ve ⟶ content Res ⟶ –ve ⟶ organisation Res ⟶ neutral ⟶ content Res ⟶ +ve ⟶ task requirement	Rater responds to either the content, language, organisation or task requirement indicated in the text. The response could be generally positive, negative or neutral.
Evaluate ⟶ content Evaluate ⟶ language Evaluate ⟶ organisation	Rater scores the text indicating bands attained.
Res ⟶ interpret ⟶ content Res ⟶ interpret ⟶ language Res ⟶ interpret ⟶ organisation	Rater interprets the content, language or organization of the text.
Res ⟶ interpret ⟶ descriptor Refer ⟶ descriptor	Rater refers to or interprets descriptor.
Res ⟶ improve	Rater suggests areas of improvement.
Res ⟶ –ve ⟶ legibility	Rater comments on legibility of text.

Example B

Except that his use of the word "emerge" is wrong throughout that he uses emerge to . . . to mean "cause" I think , but actually it takes the word "emerge" from the prompt quite literally. (InF4/10)

Res ⟶ –ve ⟶ language
(Res – Response ; –ve – negative; language – language category)

Example A shows positive rater response to the element of organization while example B shows some negative response to the element of language in the respective segments of text read. The following table presents the coding schemes used in the entire transcription.

Naturally, there were areas of overlap as certain units were analysed using more than one code as it was read as indicating more than one rater action happening as seen in example A.

Example A

Doesn't make sense. In a competitive situation the human nature, the desire of winning will emerge. So basically, this is his thesis, I think. What is his thesis? (InF2, 24)

Res ⟶ interpret ⟶ content
Res ⟶ interpret ⟶ organisation

The comment in example A was coded as both an interpretation of content ("Doesn't make sense") and organization ("What is his thesis?").

Appendix D presents samples of the transcription and coding schemes used.

Findings

Overview

This section presents results from the analysis of data gathered from the think-aloud protocol amongst the four raters. The findings are presented in four main sections corresponding to the four research questions articulated. In brief, the four sections discuss these aspects of the rating process:

(a) raters' similar practices
(b) individual/ idiosyncratic practices amongst raters
(c) areas of obscurity and phases of indecision in the use of descriptors
(d) strategies to manage indecision.

Where useful, data gathered from retrospective interviews with the raters will be cited to illustrate the respective points made.

Similar Practices amongst Raters

Raters generally follow a sequence in their rating process that involves a reading phase that precedes the actual scoring phase. In the reading phase, comments that involve management behaviours (Example A) are more frequent than in the scoring phase.

> Example A
>
> I'm going to look at 4 scripts and record my thought process as I attempt to mark them. (S100/InF1[1])

It is also at this stage where phrases related to descriptions or the actual terms within the descriptors are constantly referred to as raters make first hand impressions of the text being read (see underlined words in Example B below).

> Example B (InF2)
>
> and all spelling mistakes especially the p-u becoming p-e. And the "i" were having "e" instead and so far the syntax is so fractured and the idiom is so foreign that the meaning is practically obscured. And I'm not inclined to give a very high mark on the language. (S100/23)

Table 2 shows the average number of comments made with reference to the three major categories of Content, Organisation, and Language. No raters neglected the provision of comments on any of the major categories. At least 90% of all raters' comments concern the three major categories.

[1] InF1 denotes informant number 1 while S100 denotes the number of the script marked.

Table 2
Frequency of Comments Made on the Three Major Categories

Rater	InF 1	InF 2	InF 3	InF 4	Total
Number of comments	92	94	120	172	478
% of total comments	93%	94%	90%	97%	
Number of texts analysed	4	4	4	4	16
Average per text	23	23.5	30	43	

Comments were generally classified into negative, positive or neutral in the analysis.

Example B above shows a negative response to Language while examples C and D below show a positive response to organisation and neutral responses, respectively.

Example C (InF1)

So I think that his first paragraph has a focus thesis point. (S100/9)

Example D (InF1)

Now, I have no objection to it. (S100/15)

It is also observed that generally raters, except for rater 2, do not make any allusion to scores when they respond positively or negatively to the text in this phase. Rater 2 had two instances when some vague relation was made between the impression and score as shown in Example E below.

Example E (InF2)

This is terrible English and I'm very inclined to mark it down from the first sentence. Tenses are wrong, its illegible, the spelling mistakes, although of course it actually does address social topic and it sounds like a thesis. (S100/2)

However, the occurrence of such preliminary scoring (Lumley, 2002, p.255) was rare. Generally, the raters scored scripts only after a complete reading of the scripts although certain parts of the scripts may be read and reread before scoring takes place. Essentially, raters go through a three stage sequence similar to what Lumley (2002) describes although fewer details in each stage are captured:

(a) First reading (pre scoring) — Overall impression of global and local features
(b) Rates all three categories — Descriptor and text
(c) Considers score given — Descriptor and text

Although raters seem to go through a rather uniform sequence in the rating procedure, the consistency becomes less apparent when one investigates deeper into individual rater's interaction with the details of the descriptor and scripts in both the reading and the scoring phases.

Raters' Individual/Idiosyncratic Practices

Focus on Categories

Although raters generally paid attention to all three categories in the descriptor, it seems that some raters may at some point focus on a particular category over other categories. This is most apparent in InF2 who seems to be particularly concerned about the Language component as demonstrated by the incessant comments on details of language used in S100 in alternating comments (see Example A below).

Example A (InF2)

This is terrible English and I'm very inclined to mark it down from the first sentence. Tenses are wrong, its illegible, the spelling mistakes, although of course it actually does address social topic and it sounds like a thesis. (S100/2)

Full stop fault, the punctuation is bad plus the usual problem to the tenses and spelling. (S100/4)

Idioms is definitely off. (S100/6)

doesn't have the correct type of adjectival phrasing. (S100/8)

spelling mistake and syntactical mistake mark down. (S100/10)

more spelling mistake. (S100/17)

This example may not provide compelling evidence of rater 2's focus on the Language category as the insistence on sourcing out language errors is not as clearly apparent in all the scripts rated. However, with reference to Table 3, rater 2 seems to be relatively more frequently engaged with the Language component than the other three raters.

InF3 and 4 however seem more fixated on the Content category, especially InF4 with an average of 20.7 comments in that category. Examples B and C illustrate their comments pertaining to the Content category.

Example B (InF3)

I don't know what this means. (S100/11)

I don't know what this word is . . . as usual (S100/20)

So I'm not quite sure if its directing at, because I gets a little bit . . . a little bit confused what this writer means by however here. (S100/22)

Makes some sense here. (S100/28)

What do you mean by cover here? (S100/36)

Not sure what it means (S100/44)

Oh what is thick cover? This hasn't been explained actually. (S100/50)

Err I guess the main point is there but I am not really convinced because the main point is just been going back and forth in every paragraph. (S100/53)

Table 3
Average Number of Comments Related to Each Category

Category \ Rater	InF 1	InF 2	InF 3	InF 4	Total
Content	14.2	8	17	20.7	240
Organisation	4.5	6.5	5.7	4.5	85
Language	4.2	9	7.2	4.2	153

Table 4
Number of "Read" Entries

Rater	InF 1	InF 2	InF 3	InF 4
Read	14	33	76	138

Example C (InF4)
I guess what you mean is err . . . when someone studies better (S100/6)
Okay, I suppose when there is (S100/13)
What is he trying to achieve? (S100/17)
What does this mean, (S100/21)
Doesn't make sense. In a competitive situation . . . (S100/24)
That seems to be what he is saying. (S100/27)
What's he trying to say? Err, I don't know. (S100/28)

The tendency to pursue any one category more intensely than others affects the frequency of related rater behaviour. As in this case, InF3 and InF4, who seem more concerned about the meaning potential of points made also seem more inclined to read and reread parts of the scripts more frequently, as shown in Table 4.

Frequency of "Read" Codes

It is not unusual for InF3 and InF4 to correct students' errors as they read and reread in order to draw out the meaning potential of the texts. Such reading strategies help the raters to "develop an overall sense of how much strain the text causes on a superficial reading" (Lumley, 2005, p. 151) and this interpretation strategy helps configure an overall impression which might ultimately contribute to the scoring phase.

Use of Descriptors

Raters in this study also demonstrate different degrees of reliance on the descriptor. This is illustrated by Examples A and B below.

Example A (InF2)

Erm, we said that we like to see citations but it was not entirely correctly done. (S101/39)

We did not like the fact that he did not impact on the quotation. (S101/40)

Example B (InF3)

So in terms of content, I would, let me check, very good interpretation of the set question or a fair understanding of the set question that is no. 3. Urm urm most ideas are mostly relevant are sensible, focused ideas, main ideas are sensible. No. 4 ideas are better focused and fully developed. Fair understanding of the set question that is identifiable, insufficiently developed, some irrelevant ideas, urm (S100/54)

Example A shows InF2 referring more to an awareness of what the descriptor says, depending very much on memory for what the guidelines are. Example B, in contrast, presents a rater combing through the list rather finely to remind himself/herself of what the specifications are. The level of dependence does affect the extent of vivid details that the rater may have of the specifications and this may affect the way raters score the essays. However, such behaviours may be difficult to streamline and raters can only be reminded on the importance of being conscientious about referring to descriptors, especially in the scoring phase.

The above section presents some areas of differences in the way raters behave in the rating process. The impact that these differences may have on the scoring process or the level of variation in scores is not apparent from the present investigation. It is also not apparent if these practices are employed by the respective raters in fairly consistent ways across all the scripts that they assess. As such, to ascertain their impact requires more detailed investigation into these particular areas.

Areas of Obscurity in the Rating Process

This section presents some areas of obscurity inferred from the comments made by the raters as they interacted tenuously, at some phases, with the descriptor and the texts assessed. There were five areas of difficulties that surfaced in the analysis with three of them pertaining to the category of Content in the descriptor.

Content: Relevant Main Ideas Versus Meaning

Raters look for main ideas that are relevant as they read the scripts. The identification of these ideas is assessed under the category of Content in the descriptor. However, from the comments provided, it seems that there is a tension between the presence of main ideas in the scripts and the ease with which the meaning of the points can be decoded, i.e., relevance of content versus clarity of meaning. The examples below illustrate this tension.

Example A (InF2)

I think it sounds as if it should make sense, but I don't think it's so easy to derive sense as you have to think hard about what the author wants to say. (S101/71)

Example B (InF1)

His points are there, but . . . but he never quite make the link (organisation?) you know so, having put 4 down there sometimes try to move down and say maybe he's a 3.(S100/40)

Example C (InF3)

I think the main problem in this essay is that the ideas are pretty much unarticulated, I think the person struggles with his or her own ideas and points, in such a way that somehow I see where he or she points is going, but the organization does not capture it. (S103/161)

Example D (InF4)

Actually his ideas are there, just that he doesn't have the language to express himself. Its very hard to read, and get all very confused and have to explain it to yourself. (S100/117)

I think content should be a 4, the organization maybe its three because really hard to link the ideas. I think because I have to explain myself so much, (S100/120)

As can be seen from the examples, raters do raise comments about the ease of decoding the meaning to points made. It is interesting to note that when there is a perceived difficulty with decoding the meaning, raters may differ in their interpretation of where the lack may arise from. For instance, in example B, the rater attributes the fault to an absence of links (Organisation category) while in Example D, the rater attributes the difficulty partially to the lack in language as well as organization. The difference in perception of the source of the problem could allow variation in rating as different levels of penalty are associated with the weighted categories of Content, Organisation, and Language.

Prompt Interpretation

Another way in which essay content presents difficulty to the raters is in the interpretation of the prompts. Raters' comments show incidences of them referring to the prompt and interpreting what it requires students to write, as shown in Examples A and B below.

Example A (InF 4)

I find myself reading it and several times being convinced by it. but the writer uses "only" as a lynch pin, as I said I'm not too sure whether we should you know, encourage students to go on ONE particular word in the prompt and then develop from there. (S100/24)

Example B (InF3)

The ONLY part, I feel, is very important. (S100/115)

Okay, so this is his thesis. Alright, again this person deals with the word ONLY. So this is precisely what I said and he is addressing it. (S101/132, 148)

In the examples, both raters tried to understand the specific requirements of the prompt although the two raters seem to have contrasting interpretations of what it requires. In Example A, the rater saw the emphasis on the word "only" as signaling an unbalanced perspective while in Example B, the rater found the emphasis on the word "only" an instance of a stance which addresses the crux of what the prompt requires.

Such contrastive interpretations may have an impact on the scoring process, contributing to further variation amongst raters. This can be seen in the raters' comments in Examples C and D below.

Example C (InF1)

He has 2 views and this view is not very healthy because the way I see it, we are not encouraging students to sit on the fence erm, to sort of you know, hatch my bets, 50-50 kind of thing, in case anything goes wrong I still have 50 percent to go on. So I think that this sort of erm, you know, prove to be detrimental to this S102 script because so, he's having, his essay is going to be split up and I . . . as I read on I saw it as being that. (S102/93)

Example D (InF4)

He is trying to address some of the issue . . . on the reverse side ah.. the counter argument, and it may be true that it is his true character! Yes, that's true. Okay, he addresses the counter but now he is sticking to his stand, Okay, good! (S102/235, 237)

Okay, at least he does deal with the ONLY bit eventually, in his essay, in his paragraph, so he didn't quite mentioned it in his thesis, so . . . ya . . . he does try . . . to put in it ya . . . and he does have a contrary view, so let's see. Erm . . . (S102/243)

InF1 saw the writer's stance as "sitting on the fence" while InF4 saw it as one recognizing contrary views. As shown in Examples A and B respectively, the former signals a rather negative reaction (e.g., detrimental) while the latter signals a positive reaction (e.g., good) to the stance taken by the writer. This then is yet another area to address in the rater training sessions.

Stance

Another way in which raters may introduce a source of variation in the rating process is if they unwittingly allow their personal stance on the issue to colour their perception of what constitutes a position to take on the issue. Raters' comments show that they verbalise their agreement or disagreement with the writer's position as they grapple with the arguments presented (see Examples A and B following).

Example A (InF1)

That everybody will agree with. (S100/ 20)

Now, I have no objection to it (S100/ 15)

and that I would agree with him to a certain extent of my own feeling about that . . . (S101/88)

Example B (InF2)

Well, I really don't see how that can be true. Urm, he may . . . . (S101/49)

Well, I don't think the environment factors are excuses to take drugs. I think it has to do with integrity and moral fibre however. (S101/52)

Well, I think a lot of people could argue the other way round. (S101/60)

Its true, (S103/135)

As Examples A and B show, at certain points, raters may not be wholly agreeable to the writer's argument. These comments signaling disagreement are quite unlike comments that point out inadequacies in constructing good arguments that present strong supporting evidence. Raters should be acutely aware of the danger of biased scoring should their own stance affect their judgment of the writer's position. As InF3 explains in Example C below, raters do try to take an objective perspective and allow the students a chance to develop their own position well and they should be rewarded appropriately if they are successful.

Example C

Number one, do I often not agree with the writer? Yes. What do I do about it? I let the writer develop. If he develops it well, I will give him the point, if he doesn't he won't get the point. But that's not necessary because I disagree with him, but just because I'm waiting to see how it turns out. And it will be the same even if I agree with him. Erm, as to do I use the descriptors to decide what to do? No, because I don't think the descriptors actually tell you what to do, as far as when you disagree with something, it will just tell you to not <unclear> that the point doesn't develop very well, but it doesn't say if you disagree with point and the point is still very well develop, you can give it a high mark. (Interview data, InF3)

The final impact on marks may not be easy to trace but it is important to consider some questions on how the rating process can manage this problem. For instance, can such problems be adequately accounted for by descriptors or a chief moderator whose role is to emphasise objectivity? Also, does the writer have a more difficult job if the rater does not take the same stance, i.e., though the rater will wait to be convinced, it is an uphill task for the writer? Perhaps the interpretation of prompts is a pertinent issue that deserves more attention in rater training sessions.

Academic Tone

Besides the content of the essay, raters' comments on two other components show the lack of clarity in their use of the descriptor. One of these components involves the use of source texts in the writing of the essay. Raters do note that the use of quoted texts from the pre-reading passages is a positive practice, as shown in Examples A and B following.

> Example A (InF2)
>
> well, that sounds quite promising because he's got a citation and its looks academic. (S101/33)

> Example B (InF3)
>
> The erm, good start I guess. you know, a quote, that establishes credibility. (S101/63)

Example C, however, shows a rather negative rater response as the rater did not find the use of source text adequate.

> Example C (InF2)
>
> But he doesn't actually expand on the quotation about the creation of anxiety or hurt feelings. (S101/37)

Some issues arise when rating of these citation-related criteria is considered, as pointed out by one rater's comments in example D below.

> Example D (InF 1)
>
> I think the winning point is that he's got consistent reference to secondary sources as support. But again as I said, the descriptors doesn't erm, ask us to award you know, extra points for person who does this as . . . would be read you know, which is that you may use, the rubric says you may use if you like or if you desire or something like that (S101/80).

More fundamental to the need to streamline how or when the use of source texts should be rewarded are questions, for instance, on the level of proficiency with which undergraduates at entrance level are expected to use source texts. These questions have to be answered in the context of individual institutional requirements and standards expected of their undergraduates.

Legibility

Another area of raters' comments concerned the handwriting of the students, as shown in Examples A and B following.

> Example A (InF1)
>
> this script err . . . is difficult to read from the very beginning because in line (S100/2)

Example B (InF3)

No, idea, I really cannot stand bad handwriting. (S101/58)

It is not incredulous to suggest that legibility does affect raters but do/should descriptors take care of such items too? According to Smith (1998), factors extraneous to the performance criteria statements, such as handwriting, may be commented on but they were made by way of observation rather than by assessment and, with the exception of one rater, did not influence judgments relating to the demonstration of individual performance criteria (Smith 1998:47).

Phases of Indecision

The above section presents areas of obscurity that raters in this study encountered and this may have contributed to phases of indecision in the scoring phase, as exemplified by the comments below.

Example A (InF 1)

Erm . . . of course you know organization erm well . . . content I gave him a 3 and organization I gave him a 4 because he has no relational err patterns. His points are there, but . . . but he never quite make the link you know so, having put 4 down there sometimes try to move down and say maybe he's a 3. (S100/40)

Example B (InF2)

Urm, and for language, I think I would give it a two or even a one, because it reads really very badly. (S103/139)

Example C (InF4)

Okay, does he show recognition of topic complexity? Actually no! he got some pretty interesting ideas, so he is not quite everything above 4, 5 but maybe more still of a 4. (S101/199)

These phases of indecision could well be due to some areas of obscurity which the descriptor failed to address as discussed earlier. However, it could also well be the case of a mismatch between the vague impressions that raters have as they evaluate essays and the neatly organized performance criteria that results in immense difficulty when raters try to articulate their reasons for their respective assessments. The raters' comments below present samples of impressions that assessors make and very often, these impressions have little semblance to the categories captured in discrete segments in the descriptors.

very inconcise writing and of course, the language coming in urm urm urm, is stopping me from understanding him in a faster and more efficient way ok? (InF1/17)

he is actually quite elegant and can, still capable of writing an elegant expression and all that (InF1/29)

his writing deteriorating, seems to be more desperate (InF1/66)

Now my question here is the sense is totally incomplete. Not only is it incomplete, it's not making any sense at all! (InF1/74)

that this writer has a particular style in using transition markers well and it's excellent, (InF1/118)

I mean given the way he starts the essay you know, a very promising start and erm, however it's not matched by a consistent err, promise throughout the essay (InF1/123)

the conclusion looks good superficially, but if you think about it, he doesn't use his opportunity to expand on it and therefore as a conclusion it lacks punch

Urm, mentally structured, well structured. (InF3/97)

But somehow, I quite like this person for the kind of statements and examples, reasoning that he or she produces so far. (InF/184)

Hey that works well (InF1/7)

Paragraph four, urm, he uses interesting words, but they are not well used, not clearly used and not elaborated on.(InF4/127)

As can be seen from the underlined portions of the examples, the articulation of impression gained may have little relation to the clear and discrete description of language use in the descriptor. Lumley (2005, p. 241) points out the tension between reliability and the "disordered impression" the rater may have gained of the text which ultimately translates into "a tension between the publicly accessible and visible scale descriptors and the rater's privately inaccessible and intuitive impression." According to him, " . . . from the rater's point of view, the articulation of scores is close to impossible, because it relies on a deeply internal and inaccessible but general impression of the text" (Lumley 2005, p. 206). Lumley therefore argues that ". . . the main function of descriptors and training is to help raters channel their diverse reactions to texts into narrower, more manageable statements that meet institutional requirements" (Lumley, 2005, p. 255).

Strategies to Resolve Indecision

In the rating process, the rater ultimately has to make a decision in reconciling descriptor and text and decide on a score. Indecision at this stage of the process is exemplified in Examples A and B below:

Example A (InF1)

. . . and organization I gave him a 4 because he has no relational err patterns. His points are there, but . . . but he never quite make the link you know so, having put 4 down there sometimes try to move down and say maybe he's a 3. (40)

Example B (InF2)

Urm, and for language, I think I would give it a two or even a one, because it reads really very badly. (139)

Generally, raters resort to some form of strategies to reconcile discrepancies in the script and descriptor so that a score can be arrived at. One strategy that is used is for the rater to comb through the descriptor very finely and then to somehow settle on one factor to tip the scale towards a certain band. This is shown in Examples C and D below.

Example C

I think I refer to the descriptors when I find myself really trying to, how should I say it, quantify, when I'm trying to rate the essay, meaning that when I read the essay, I tend not to refer to the descriptor too closely. (Interview data, InF4)

Example D (InF4)

Urm I would probably go for four. Urm . . . number 3 says . . . its really irrelevant ideas. Yes, I think number three, more of number three. Its really quite rambling, makes sense but there was a thesis. There was a clear thesis statement to me but I would like to go for a three this time because of the "rambling" that has been going on here. I think I will definitely, most probably go for a three. (S100/55-59)

In Example D, InF4 saw "rambling" as severe enough to place the scoring at a particular band.

Another rater relied on discussion points raised by other colleagues in the training session to help guide her scoring (see Example E below).

Example E (InF1)

Then I always remembered that when I always co-marked with other people, and you know, like comments thrown out and all that, I do hear people giving people second chances, and then they say things like, ... (Interview data)

As Pula and Huot (1993) explain, raters take their assessment criteria from their teaching experience, not their experience as readers.

At some point, it is the impressionistic notion of the degree to which a candidate has fulfilled certain performance criteria that helps the rater arrive at a score (see underlined expressions in Example G. This is exemplified in Examples F and G below.

Example F (InF2)

Urm, so I would give him perhaps 3 on the content. But I would really like to downgrade it to two, but I don't think it's entirely a two, because it's not entirely superficial (139).

Example G (InF4)

Good interpretation, fairly good interpretation, I suppose it's about the same as the first one, but it says that some of his ideas are a bit more sensible interesting, er . . . main ideas are interesting and well developed, shown recognition of topic complexity, actually he doesn't quite show recognition topic complexity but, he has done quite a good job of showing some of the, or even countering some of the arguments, you know, sensible, interesting ideas, can still be better focus and develop most, ideas are mostly relevant. So what's the difference? Okay, does he show recognition of topic complexity? Actually no! he got some pretty interesting ideas, so he is not quite everything above 4, 5 but maybe more still of a 4. (197-199)

The descriptor is constructed in such a manner that there is a gradation of various criteria that differentiates the bands. In many instances, it is difficult to describe the presence or absence of these performance criteria in objective or quantitative terms. However, such qualitative descriptions allow a level of subjectivity to seep in. As such, raters have to be conscious of this source of variation which probably can be minimized with repeated experience in rating similar scripts and deliberating on the scores associated with these scripts of respective levels.

As can be seen from the above account, the rating process is much more complex in reality as compared to the clear categories reflected in the descriptors. As Endrosy (2000, p. 113-114) points out, rating scales are not the sole determinants of writing quality in raters' judgment, but are regarded by raters as only one (even possibly the weightiest) of several factors that they must take into consideration.

Discussion and Implications

The findings above present insights into how raters operate or how they report they operate. In the pre-scoring stage, the raters' rather elaborate impression of the quality of text is important although no actual scoring is done. The factors that contribute to the raters' intuition at this stage are unclear although it may be the fundamental step in assessment. In the scoring stage, the fundamental role of the descriptor in channeling raters' responses into manageable and discrete categories is clear but there are many other avenues that raters fall back on besides the descriptor. There seems to be no straightforward relationship between the descriptor and the rating decision in problematic areas. Strategies used may/may not be those approved in the training process but ultimately, raters try to fit impressions of texts to the descriptor even when it is hard to do so.

According to Cumming et al. (2001), there is increasing agreement about the range of features that fall into the categories that are pertinent as performance criteria. The agreement is based on the broadly common experience of raters gained from extensive language teaching and testing. However, as the study shows, there may be features that need to be modified or categories that need to be extended to provide an even more accurate assessment tool that will

reflect prominent elements that raters tend to look for in essay assessment. The differentiation between the presence of a point and the difficulty involved in the decoding of a point is a good example surfaced in this study. Perhaps, the feature of 'clarity of meaning' is an item that needs clearer representation in the descriptor as raters frequently comment on the difficulty in decoding points.

The proper use of sources is another case in point. If the use of sources is a feature in task fulfillment, then the descriptor must address and incorporate description of relative levels of satisfactory performance in clear terms. Another area that needs attention is the way in which a prompt is interpreted. As seen, the interpretation of a prompt affects that raters' idea of what constitutes relevant points and more generally, what constitutes a good argument. As such, the effect is pervasive enough to warrant closer attention so that raters come to a consensus on the requirement of the prompt. Thorough preparation in selecting good samples of a range of scripts that represent various possible atypical and typical student responses would be necessary for a successful training session. In some situations, tight deadlines in the rating process of large-scale tests may prevent the thorough search for good samples of representative scripts. For an area such as this, and possibly also for other aspects of a descriptor, raters are aware that there are too many possibilities for the scale to capture them all, and even if it did, there would be an unwieldy set of conflicts for raters to deal with (Lumley, 2005, p. 296). These possibilities may be better dealt with during a rater training session than by a static descriptor. The possibilities discussed may surface other related issues and raters involved can then deliberate on common courses of action. Such sessions will also benefit from the presence of a 'chief examiner' who will moderate extreme positions and extend rater responses that are too narrow. In short, the chief examiner can harness disparate views and responses and where it is necessary, insist that certain lines be towed while at the same time identify areas where there can be some margin for varied responses.

Other factors that raters commented on, including handwriting, may or may not affect the rating of the essay but if the item is frequently raised by raters, it indicates a need to address the possibility of incorporating such factors at least in the training process so that raters have an avenue to reconsider the importance of these factors.

The training process is also an opportune time for raters to be reminded of individualistic tendencies that should be avoided such as the imposition of the raters' own belief about certain issues in the marking process. Raters should try as far as possible to remain objective and withhold judgment to allow the student to develop his/her position. Other reminders could involve the need to put as much emphasis on the different categories as the weighting of categories allow for. At certain junctures, it may be good to remind raters to read the essays well before forming an impression as it can be seen that raters do differ in their frequency of 'read' category. Essentially, the training process is a good platform for highlighting various issues in the rating process that are best captured by deliberation and negotiation rather than by a static document. These issues differ according to various factors including institutional emphasis, the nature of the prompt, and the raters involved.

At some points, decisions might have to be made about the suitability of raters for the task, especially if some raters are repeatedly unresponsive to important reminders and can be identified as consistently contributing to varied scoring. However, the selection of best raters is a luxury that may not be affordable in all rating situations as there are considerations of the number of raters needed for a large-scale test and the amount of funds allocated for such purposes if the raters are to be paid.

If resources allow, it is also a good strategy to do multiple ratings where each script is rated by more than one rater and where there is a clear procedure for reconciliation of varied scores. However, such strategies are again limited by manpower availability and time constraints.

Though the refinement of descriptors and training workshops are vital to rating consistency, McNamara (1996, p.118) points out that ". . . even with proper training, substantial differences between raters will persist with important (and unintended) possible consequences for the candidate . . . rater differences are reduced by training but do persist." McNamara thus recommends a measurement procedure using the multi-faceted Rasch model.

According to McNamara (1996, p.133), "the model states that the likelihood of a particular rating on an item from a particular rater for a particular candidate can be predicted mathematically from the ability of the candidate, the difficulty of the item and the severity of the rater." Essentially, the model allows one to say rather accurately the sort of challenge that a candidate had to face on a particular item with a particular rater and therefore this facilitates the interpretation of the actual rating given. It is also able to provide insights into the relative severity level of raters having taken into account their scoring patterns displayed amongst the total cohort of students attempting questions of different difficulty levels. McNamara (1996, p. 118) claims that the model ". . . has the potential to illuminate these issues very clearly, and to allow us to control the rating process better than we have been able to in the past." This would then be a good area for further research as rater behavior in relation to the descriptors can be described quantitatively on a linear scale to facilitate more accurate testing and marking on a large scale.

Conclusion

The awareness of inconsistency in inter-rater reliability has resulted in the adoption of various practices such as rater training and double marking of scripts in the assessment of writing. However, a detailed understanding of the specific factors that contribute to inconsistency in assessing writing will provide insights that will result in more concrete measures to manage potential areas of lack within the context of individual institution's requirements and desired standards in assessment. This study represents one such investigation that provides detailed insights into particular areas in the descriptor that need further work and refinement. Issues surfaced such as the need for finer differentiation in categories, or the addition of other categories. Other insights show the need for clear reminders and the central role that a chief moderator can play to

ensure more consistent practices. Ultimately, though inter-rater inconsistency in assessment cannot be eliminated, this investigation represents one more step towards minimizing it.

ACKNOWLEDGEMENTS

The project was funded by the Research Fellowship awarded by SEAMEO Regional Language Centre. I would like to thank SEAMEO RELC for the project funding that made this paper possible.

THE AUTHOR

Wu Siew Mei lectures at the Centre for English Language Communication, National University of Singapore. She has taught English language and communication courses both in Melbourne and Singapore. Her research interest and publications mainly involve the areas of academic discourse, writing assessment, and evaluative expressions in academic writing.

Correspondence concerning this article should be directed to A/P Wu Siew Mei, Centre for English Language Communication, National University of Singapore, 10 Architecture Drive, Singapore 117511; e-mail: elchead@nus.edu.sg.

References

Bauer, B.A. (1981). A study of the reliabilities and cost-efficiencies of three methods of assessment for writing ability (ERIC Documentation Reproduction Service No. ED 216357).

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

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*, 65-81.

Connor, U., & Carrell, P. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J.G. Carson & I. Leki (Eds.), *Reading in the composition classroom* (pp. 141-160). Boston, MA: Heinle and Heinle.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning, 39*, 81-141.

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

Cumming, A. (1997). The testing of writing in a second language. In C. Clapham & D. Corson, (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 51-65). Dordrecht, Netherlands: Kluwer.

Cumming, A, Kantor. R., & Powers, D.E. (2001). Scoring TOFEL essays and TOFEL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework. *TOFEL Monograph Series*, MS-22. Princeton, NJ: ETS.

Edrosy, M.U. (2000). *Exploring the establishment of scoring criteria for writing ability in a second language: The influence of background factors on variability in the decision-making process of four experienced raters of ESL compositions*. Unpublished MA Thesis, OISE, University of Toronto.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Green, A.J.K. (1997). *Verbal protocol analysis in language teaching research*. Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In B. Kroll (Ed.) *Second language writing* (pp. 69-87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). Scoring procedures. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Homburg, T. (1984). Holisitic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18*, 87-107.

Huot, B.A. (1988). *The validity of holistic scoring: A comparison of talk-aloud protocols of expert and novice holistic raters*. Unpublished PhD Thesis, Indiana University of Pennsylvania.

Huot, B.A. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*, 201-213.

Huot, B.A. (1993). The influence of holistic scoring on reading and rating student essay. In M. Williamson & B.A. Huot (Eds.) *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219-240.

Linacre, J.M. (1989). *Many faceted Rasch measurement*. Chicago, IL: MESA Press.

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

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Peter Lang: Frankfurt am Main.

McNamara, T. (1996). *Measuring second language performance*. Longman: London.

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquim, Cambridge and Arnhem* (pp. 92-114). Cambridge: Cambridge University Press and University of Cambridge local Examinations Syndicate.

Sakyi, A.A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A.J. Kunnan (Ed.), *Fairness and validation in language assessment. Selected papers from the 19th Language Testing Research Colloquim* (pp. 129-152). Orlando, Florida.

Sakyi, A.A. (2001). Validation of holistic scoring for ESL writing assessment: A study of how raters evaluate ESL compositions on a holistic scale. In A.J. Kunnan (Ed.), *Fairness and validation in language assessment*. (pp. 130-153). Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Sparks, J. (1988). Using objective measures of attained writing proficiency to discriminate among holistic evaluations. *TESOL Journal, 9*, 35-49.

Stratman, J.F., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: Issues for research. In P. Smagorinsky (Ed.), *Potential problems and problematic potentials of using talk about writing as data about writing processes*. (pp. 89-114).

Taylor, L.B. (2004). Current issues in English language testing research. *TESOL Quarterly, 38*, 1.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex.

Vaughan, C. (1992). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-26). Norwood, NJ: Ablex.

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

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

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

Appendix A

Sample Profile Band Descriptors

BAND	Content—ideas, arguments & evidence	Organisation—communcative quality, coherence & cohesion	Language—vocabulary, grammar & sentence structure
6	• excellent interpretation of the set Q • main and supporting ideas are extremely original, interesting, relevant and excellently and fully developed, demonstrating maturity in handling the topic's complexity	• focused introduction with an excellent thesis statement • ideas are very clearly organised with an extremely clear relational pattern (e.g. comparison/ contrast, sequence, cause/effect, order of importance, etc.) • conclusion addresses the thesis excellently with much thought and is in sync with the rest of the essay • extremely cohesive—excellent use of transition elements	• excellent sentence variety—excellent blend of simple, compound & complex sentences • extremely fluent & very sophisticated • excellent vocabulary & word choice with very accurate use of idiomatic expressions • almost no grammar, punctuation and spelling errors
5	• good interpretation of the set Q • main and supporting ideas are interesting, relevant and well developed, showing recognition of the topic's complexity	• focused introduction with good thesis statement • ideas are well organised with a clear relational pattern • conclusion addresses the thesis fully and is in sync with the rest of the essay • very cohesive—good use of transition elements (connections are generally successful with minor problems only)	• good sentence variety —good blend of simple, compound & complex sentences • highly fluent & fairly sophisticated • good vocabulary & word choice with flexible use of idiomatic expressions • few grammar, punctuation and spelling errors
4	• fairly good interpretation of the set Q • main ideas are sensible & interesting but ideas can still be better focused and developed • ideas are mostly relevant	• fairly focused introduction with clear thesis statement • ideas are fairly well organised with a relational pattern but they could be more effectively explained at the macro, paragraph and sentence levels • conclusion addresses the thesis partially but is still in sync with the rest of the essay • cohesive — fairly good use of transition elements (connections are not always successful)	• fairly good sentence variety—fairly good blend of simple, compound & complex sentences • fairly fluent • fairly good vocabulary & word choice with some idiomatic expressions inaccurately used • some grammar, punctuation and spelling errors which occasionally obscure intended meaning

Appendix B

Instructions to Raters

When you mark the scripts in this round, it is very important for you to think your thoughts aloud as you go through the essays. As soon as you have identified the script number on the top left hand corner, you can vocalize all your thoughts on the essay just like you did in the training sessions. At certain points, I may prompt you to keep talking if you should lapse into a period of silence. Otherwise, I will just be mostly listening and nodding in agreement to your think aloud comments.

Appendix C

Examples of Interview Questions

- Do you find yourself using the descriptors in a sequential way? If so, briefly describe the sequence.
- What are some general areas of difficulties that you have experienced in the use of the descriptors?
- What do you do when you recognize that the descriptors do not offer satisfactory descriptions that match your assessment of some aspects of the essay?
- To what extent do you gauge yourself as having used the descriptors fully/sparsely? What is your rationale for doing so?

Appendix D

Sample Think-aloud Protocol (Informant 1)

1. Erm.. Think allow protocol and its done by PC on the 22nd of Dec 2005. Erm.. in SELF. I'm going to look at 4 scripts and record my thought process as I attempt to mark them. Script S100. err.. //	Manage → procedure
1. this script err.. is difficult to read from the very beginning because in line //	Res → –ve → legibility
1. I'm trying to figure out whether it's a title or it continues, but err.. I sort of figure out it's a title //	Res → interpret → content
1. and it's a very strange title anyway, //	Res → –ve → content
1. not only it is non-standard and ungrammatical, its err.. it's incomplete. //	Res → –ve → language
1. Now it's her first line runs with actually starts with "we run faster than when someone is running around" //	Read
1. and I thought that: Hey that works well//	Res → +ve → content
1. and he/she is trying to show that in competitive situation people work harder. //	Res → interpret → content
1. So I think that his first paragraph has a focus thesis point //	Res → +ve → organization
1. except that his use of the word 'emerge' is wrong throughout that he uses emerge to.. to mean 'cause' I think , but actually it takes the word emerge from the prompt quite literally.//	Res → –ve → language (+ reason)
1. Now paragraph 2 is also proving to be difficult because I don't understand what he means by //	Res → –ve → legibility (+ reasons)
1. 'Potential' would be meaningless word if everyone can use whenever they want! //	Read
1. I always thought that 'potential' is something that is buried in it you know, //	Refer → background knowledge
1. but I suppose what he's trying to say is that if potential is there but we don't use it all the time but it comes out when one is pushed to it in a competition and the whole of the first half of paragraph 2 is his he attempt to illustrate the, the word emerge because he says abilities hide deep in one's body and mind.//	Res → interpret → content
1. Now, I have no objection to it //	Res → neutral → content

Appendix D (continued)

1. except that when I'm going through the 2nd paragraph, I'm thinking a lot this question: where is all these leading to? What is he saying? I know in a way he's saying that when people around you are trying very hard, you know you have no choice but to try hard as well, so I think that I understand,//	Res → interpret → content
1. but it's just very inconcise writing and of course, the language coming in urm urm urm, is stopping me from understanding him in a faster and more efficient way ok? Of course paragraph 2 is very long and inconsise, so he goes on and talks about err.. and stuff you know?//	Res → –ve → language (+ reason)
1. But the first significant point that he makes in paragraph 2 appears on page 2//	Res → +ve → content (+ reason)
1. when he says "that's why only in Olympic do you have the err.. the best athletes competing and therefore the records."//	Read
1. That everybody will agree with.//	Res → +ve → content
1. So I think that in paragraph 2 he has that last point which ends it, making a very large impact. Then I know, understand that he's potential is now increasing, he's getting a bit better //	Res → +ve → content (+ reason)
1. and I read paragraph 3. //	Manage → Procedure
1. Then again he says we can never tell a man he's good or bad.//	Read
1. But what is it, how is it related?//	Res → interpret → content
1. So I think what this candidate lacks is the connector, how is paragraph 2 connected to paragraph 3, is it an elaboration? Is it another way? Is it another example? He doesn't quite tells it but continues to goes on in he own merry way, //	Res → –ve → Organisation (+ reason)
1. and of course, again the grammar 'staff' for stuff you know, sort of gets in.//	Res → –ve → language
1. Now, the moment that I think he's not capable of doing better or writing better, he surprises me.//	Res → +ve → content
1. He says "he rose upon in chaos, devils also dont show up in usual situation."//	Read
1. and I say that well, he is actually quite elegant and can, still capable of writing an elegant expression and all that.//	Res → +ve → language

Appendix D (continued)

1. Alright, so I don't give up hope and I read paragraph 3 hoping that things becomes better,//	Manage → Procedure
1. and then, he goes on saying that "oh well, you know a person true self can show up if surrounded by competitors, and his good mind though, he may be defeated." //	Read
1. Although a person loses, his character actually surfaces, and I understand that okay //	Res → +ve → content
1. erm.. and I think that he.. he has better potential to say more things but he doesn't because he ends like that.//	Res → –ve → content
1. He says "competition helps us performance again.",	Read
1. problem with words form and all that. //	Res → –ve → language
1. So in general, these guys have 1 or 2 good point,//	Res → +ve → content
1. which would be clearer to the reader and markers if he had only made connections.//	Res → improve → content
1. You know, his transition markers is intensively poor, so for this candidate,//	Res → –ve → language
1. for content, what I can say is that one may identify this candidate's 2 or 3 main points, okay, but erm.. one tries very hard in order to do that. One has to sort of take a look at okay.. he's saying that, he's saying that, okay what is the paragraph connection. So you make a lot, you do a lot of work on your own erm.. for this candidate, and I think that he could have developed his cases and examples a little bit more, because we have 2 things working against him – he doesn't write well, he has poor erm.. use of language, he doesn't know how to use language. He's has lots of none standard uses of language and grammar. He has erm.. very few good examples and cases that stand your mind.//	Res → –ve → content/language
1. Erm.. of course you know organization erm well.. content I gave him a 3 and organization I gave him a 4 because he has no relational err patterns. His points are there, but.. but he never quite make the link you know so, having put 4 down there sometimes try to move down and say maybe he's a 3.//	Evaluate → organisation (+ reason) Evaluate → content (+ reason)

Appendix D (continued)

1. Of course his language is intensely very poor, you know and erm.. he's got problems with word form, erm performance for perform, he used the word "emerged" wrongly, in all 4, 5 times in the essay, staff or stuff. Prepositions are missing, there's pronoun disagreement, active confused with passive voice and loads of spelling mistake. So I think that this has got a lot against him, this.. this writer okay.//	Res → –ve → language
1. And additionally if one were to quibble with.. with erm this particular writer, he's made no reference to, he's made no reference to secondary sources at all. //	Res → –ve → task requirement
1. Now of course the rubric says that you may or may not use the sources so you know//	Refer → descriptor
1. one can't take it against him. We really can't take it against him//	Evaluate → content
1. but he seems to be going on his own strength and he doesn't really have a much.. very much on his own to stand with //	Res → +ve → content
1. so one would have thought that he would have want to use the secondary sources to at least back him up, to get more points you know, and stuff like that.//	Res → improve → content
1. Okay, so that is S100 huh, erm which I find an incredibly difficult err.. essay to read, not only because the transition points are weak but because the.. the language I've consistently got to say err it's not this it's that, he means this and he means that, so that's taking a lot of time.//	Res → –ve → language (+ reasons) Res → –ve → organization (+ reasons)
1. Erm, S100 starts very differently.//	Res → compare
1. erm starts with a secondary source so erm I find that a very strong start because it really you know, gets very well, it's very related to what the prompt asked for.//	Res → +ve → content/task requirement
1. but the moment I say that, I read his quotation from Nelson//	Manage → Procedure
1. and I discover that he is taking a part from the secondary source that would help him because later as I discover and I read the essay and I come back to it again I say I know why because he starts off with cheating. He starts off with Nelson's erm.. quotation that says//	Res → +ve → content
1. it leads to cheating and it hurt feelings, okay and it leads to hurt feelings as well. //	Read

Appendix D (continued)

1. So I said hmm not too bad you know.. erm in some kind of an attempted thesis	Res → +ve → Organisation
1. but although that's not the point okay. and that is confirmed at the end of the paragraph 1, at the end of his thesis //	Res → –ve → Organisation
1. but he says "in such a detrimental position err, situation, a person's real potential and character cannot be accurately and fairly assessed"//	Read
1. but the prompt does not ask for you to assess, okay?//	Res → interpret → task requirement
1. So, I said never mind I'll give him a chance see whether develop said.//	Manage → Procedure
1. Now, if I accept him, and which I did, I.. I find that I had.. had to accept his thesis at this point of time, then the rest of it would make sense because paragraph 2 starts with "competition may lead and otherwise morally sound person to cheat."//	Res → interpret → content
1. Obviously I'm not surprised because he started with "cheat" in paragraph 1 and therefore he would want to.. to follow with okay?//	Res → +ve → organisation
1. And then I read on and he talks about cheating and erm.. and.. a lot of.. a lot of talk goes on cheating //	Res → interpret → content
1. so I thought: opps! Here is this person going on one particular sub-aspect or sub-system you know, and he goes on to talk about it". //	Res → –ve → content
1. And I find myself asking the question over and over again "where is all these leading to?" Okay, it's cheating, I know it's fine, it's no good, you know, it's part of competition and cheating is not where a person's character will shine okay.//	Res → interpret → content
1. Now, I.. I err have this.. this suspicion of mind that he's going off in a different tangent is confirmed again in the end of paragraph 2//	Res → –ve → content
1. when he says "therefore, a competition is not a good situation to measure a person's true character."//	Read
1. then I said hmm.. again he's harping on measuring. so I think, here's this person who's dead set on the word on competition measuring a person's potential and character when the prompt doesn't ask for that really.//	Res → –ve → content/task requirement

Appendix D (continued)

1. Err.. when it comes to paragraph 3, I find his writing deteriorating, seems to be more desperate okay, and.. and this is shown by his erm, use of suspicious, suspicious thoughts 3 times in the paragraph.	Res → –ve → content
1. So I said, what is all this, what's happening here and all that. //	Res → interpret → content
1. So, erm, I read on //	Manage → Procedure
1. and I find that paragraph 3 doesn't really have very much because it's about how.. how bad, how bad a person gets during competition, what it forces him to the ultimate okay, not just cheating, but as being suspicious of his competitors and that's bad for him.//	Res → –ve → content
1. Then finally in P4, erm.. erm I think I get a confirmation, final confirmation that this person is actually, seems to be answering a prompt that says "what's so bad about competitions?", "what are the demerits of competition?" if the prompt had been that, he would have do very well!//	Res → –ve → content/task requirement
1. But here is him going off on a different slant, and I think the slant is a very small sub-set of what the prompt expects writers to write.//	Res → –ve → content
1. So the whole of P3 I have asked, I have myself asking what is the point of all these? What is the sense of him saying 'what is the point trying so hard at a competition'//	Res → interpret → content
1. and at the end saying at paragraph 4, erm, it is more realistic that this miracle sports person joins the Olympics for the experience.//	Read
1. Now my question here is the sense is totally incomplete. Not only is it incomplete, it's not making any sense at all! So I have many senses at the margin and stuff like that. Okay so, erm, it is not a good method.//	Res → –ve → content
1. At paragraph 5 he says.. he ends the essay to say it is not a good method to find a person's true potential and character. Now, I asked myself again, I tell myself again the prompt actually doesn't ask for you, for the writer to prove that is a good method because the prompt doesn't ask for it as a good method, okay? It is ludicrous to even suggest that, you know, competitions be a method of measuring people's true character.//	Res → interpret → content/task requirement