Assessment criteria in a large-scale writing test: what do they really mean to the raters?

Tom Lumley

Language Testing 2002; 19; 246
DOI: 10.1191/0265532202lt230oa

Assessment criteria in a large-scale writing test: what do they really mean to the raters?

Tom Lumley *Hong Kong Polytechnic University*

The process of rating written language performance is still not well understood, despite a body of work investigating this issue over the last decade or so (e.g., Cumming, 1990; Huot, 1990; Vaughan, 1991; Weigle, 1994a; Milanovic *et al.*, 1996). The purpose of this study is to investigate the process by which raters of texts written by ESL learners make their scoring decisions using an analytic rating scale designed for multiple test forms. The context is the Special Test of English Proficiency (*step*), which is used by the Australian government to assist in immigration decisions. Four trained, experienced and reliable *step* raters took part in the study, providing scores for two sets of 24 texts. The first set was scored as in an operational rating session. Raters then provided think-aloud protocols describing the rating process as they rated the second set. A coding scheme developed to describe the think-aloud data allowed analysis of the sequence of rating, the interpretations the raters made of the scoring categories in the analytic rating scale, and the difficulties raters faced in rating.

Data show that although raters follow a fundamentally similar rating process in three stages, the relationship between scale contents and text quality remains obscure. The study demonstrates that the task raters face is to reconcile their impression of the text, the specific features of the text, and the wordings of the rating scale, thereby producing a set of scores. The rules and the scale do not cover all eventualities, forcing the raters to develop various strategies to help them cope with problematic aspects of the rating process. In doing this they try to remain close to the scale, but are also heavily influenced by the complex intuitive impression of the text obtained when they first read it. This sets up a tension between the rules and the intuitive impression, which raters resolve by what is ultimately a somewhat indeterminate process. In spite of this tension and indeterminacy, rating can succeed in yielding consistent scores provided raters are supported by adequate training, with additional guidelines to assist them in dealing with problems. Rating requires such constraining procedures to produce reliable measurement.

Address for correspondence: Tom Lumley, Asian Centre for Language Assessment Research, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong; email: egluml@polyu.edu.hk

 10.1191/0265532202lt230oa

I Introduction

The use of performance assessment as a measure of writing proficiency in English as a mother tongue as well as English as a Second Language (ESL) is so widespread that it is rarely questioned these days, although there was a period when it fell out of popularity in the USA, as Spolsky (1995) recounts in detail. A major reason for this was the lack of reliability achievable in assessment based on ratings, which are necessarily subjective. The reliability of writing performance assessment has been improved over the years through a combination of:

- training (McIntyre, 1993; Weigle, 1994a; 1994b);
- better specification of scoring criteria (Jacobs *et al*., 1981; Alderson, 1991; Hamp-Lyons, 1991; Weigle, 1994a; North, 1995; North and Schneider, 1998); and perhaps, to some extent,
- tasks (Hamp-Lyons, 1990; 1996; Kroll and Reid, 1994).

While the pursuit of reliability remains an essential consideration researchers have also pointed out, over a period of time, how the validity of performance assessment has been insufficiently addressed. As Cumming *et al*. (2001: 3) comment: 'a principal criticism has been that the exact nature of the construct they assess remains uncertain.' Moss (1994; 1996) has also questioned the role of reliability as a necessary but insufficient condition for validity in the context of performance assessment.

An important aspect of investigating both validity and reliability is concerned with how the process of rating is managed. Huot's position (1990: 258), referring to rating of first language (L1) writing, was unequivocal:

> Other than results that measure the importance of content and organisation in rater judgment of writing quality, 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.

More recently he has repeated his call for further investigation of the rating process (Huot, 1993; Pula and Huot, 1993), while Vaughan (1991), Hamp-Lyons (1991), Weigle (1994a; 1994b; 1998), Connor-Linton (1995), Cumming (1997) and Kroll (1998) have all made a case recently for further exploration in this area in the context of ESL assessment.

Concerns focus on issues such as the superficiality of rating scales in comparison with the complexity of written texts and the readings made of them. Charney (1984), in an influential article, raised a number of questions about holistic rating. She recognized a range of factors as relevant to the rating process. She categorized certain factors

as necessary to improving reliability. These include those associated with training (including peer pressure, monitoring and rating speed), topic type and choice of criteria. She also hypothesized that other factors will undermine reliability, in particular, raters' idiosyncratic criteria, tiredness and thoughtful response. Essentially, Charney insists that the rating must be both quick and superficial in order for it to be reliable, and that deeper consideration of the text, or thoughtful response, must lead to reduced reliability. In contrast, Huot (1993) found evidence that this is not necessarily the case.

Cumming *et al.* (2001: 3) view the problem in this way:

> The simplicity of the holistic scoring method, and the rating scales that typically accompany it, obscures its principal virtue: reliance on the complex, richly informed judgements of skilled human raters to interpret the value and worth of students' writing abilities.

In response to calls for investigation of rating behaviour, studies over the last decade or so have focused interest on the rating process. In second-language assessment contexts, the major studies include Cumming, 1990; Vaughan, 1991; Weigle, 1994a; 1994b; Milanovic *et al.*, 1996; Zhang, 1998; Cumming *et al.*, 2001; Lumley, 2000. These studies have taken the approach of consulting raters directly and requiring them to describe the rating process in verbal reports produced as they rate, essentially following methodology described by Ericsson and Simon (1993).

One focus in such studies (e.g., Cumming, 1990; Cumming *et al.*, 2001) has been to identify the criteria that raters use to select and describe themselves while they are rating, rather than to examine how raters apply specified scales. The recent study by Cumming *et al.* suggests that there is now a degree of emerging consensus in studies conducted since 1990 over the features that raters refer to while rating. The present study builds upon Cumming (1990), who found that of the 28 categories of comment that he identified, 20 could be classified under the three categories (or dimensions) for which raters had to award scores: substantive content, language use and rhetorical organization. The remainder he classified as aspects of a 'self-control focus' used by the raters. However, this line of research does not shed light on the issue of how raters apply rating scales they are given to work with.

Other research has provided evidence that different groups of raters behave differently in various ways (e.g., Zhang 1998), as well as that training and experience improve agreement amongst raters. Weigle (1994a; 1994b) and Shohamy *et al.* (1992) suggest that training plays an important role in influencing raters' behaviour, especially by clarifying rating criteria. Weigle (1994a) found that rater reliability increased as a result of training, and that the improved agreement was

the result of raters gaining better consensual understanding of the terms and levels represented in the scale 'even though a host of other factors seem to be involved in the rating process' (p. 204). Levels of reliability are relatively easy to calculate. What is less clear is what the basis of the ratings actually is: how can we account for this host of other factors?

Collectively the studies referred to above form a research initiative focusing on the decisions and criteria used by raters which – as Cumming (1997) has pointed out – has the general goal of contributing to our understanding of the construct of second language writing. However, despite this body of work, the investigation of the rating process and the basis of raters' decisions is still at a preliminary stage. The present study deals with the use of rating scales by raters in a large scale assessment context. Some of the important questions that require further examination are concerned with what raters actually do with the scoring categories they consider, in particular the extent to which the raters act in similar ways to each other, and whether or not these behaviours are likely to influence rating outcomes.

With these considerations in mind, the two research questions addressed in this article are:

- Does a group of experienced raters understand and apply features of a rating scale in ways that are similar to each other?
- What does this tell us about the role of the rating scale?

II The context for the study: the *step* test

The context for the study was the Special Test of English Proficiency, the *step*, a high-stakes test administered on behalf of the Australian government as part of the immigration process (Hawthorne, 1996; Lumley, 2000). In the late 1980s, encouraged by government policy, there was a dramatic expansion in the number of fee-paying international students arriving in Australia. This was especially true of students entering the English Language Intensive Courses for Overseas Students (ELICOS) sector and, amongst these, of students from the People's Republic of China (PRC). The great majority of PRC students arriving between 1987 and June 1989 remained in Australia as immigrants (Hawthorne, 1996), where in many cases they were later joined by dependants. Following the events in Tiananmen Square in June 1989 the rate of arrivals from PRC declined somewhat but, nevertheless, between 4 June 1989 and October 1992 (when entry policy was tightened in response to the high rate of 'overstaying'), many thousands more PRC students continued to arrive. A large proportion of this latter group indicated their desire to settle in Australia

by claiming refugee status. In addition to these asylum seekers, there were substantial numbers of others, who had been resident in Australia for varying periods, from countries including the former Yugoslavia, Pakistan and Sri Lanka, as well as small numbers from other countries.

Faced with hugely inflated numbers of refugee applications, the great majority of whom had arrived from the PRC in the period following the Tiananmen Square massacre, by late 1992 the Australian Department of Immigration and Ethnic Affairs (DIEA) was anxious to resolve the status of this very large group of temporary residents. In many respects they represented ideal immigrants: for the most part they were young, relatively well educated, and already well settled in Australia. Partly because of the enormous costs and time involved in processing their claims in the courts, the government proposed a pragmatic solution, which would nevertheless be politically defensible (to avoid accusations of queue-jumping), of allowing these asylum seekers to apply for a new class of visa issued as an extension to the skilled migration program.

On 1 November 1993, the DIEA announced a special one-off visa category, the Special Permanent Entry Permit Class 816, for applicants for asylum in Australia. This created the opportunity, though not the obligation, for eligible individuals to apply for this visa. Those who chose not to go through this process could continue to be assessed as political refugees, although recent experience suggested that their chances of being granted refugee status were not high (Hawthorne, 1996: 18–19). Class 816 visa applicants were to be assessed by criteria routinely used for prospective immigrants applying from overseas under the skilled migration program. In addition to satisfying criteria of youth (aged under 45), good health and character, they would be required to pass an English language test, the *step*, thereby showing their ability to 'meet English language and post-secondary education standards' (DIEA 1995), at what was termed 'functional' level. Hawthorne (1996) has discussed the political role of language tests, including the *step*, in immigration decisions in Australia. Clearly, the use of language tests as political instruments highlights the issue of consequential validity (Messick, 1989; Shohamy, 2001), including the need for empirical investigations of the consequences of test use for test-takers and the community, especially in high-stakes assessment situations such as the one outlined here.

The *step* was announced in November 1993 and first administered in November 1994, allowing all test-takers a minimum of 12 months to prepare for the test; over the ensuing two years over 12 000 candidates sat the test. Provision was made for the minority failing part or

all of the test automatically to retake those parts they had not passed, and eventually the vast majority of candidates (around 90%) passed the test. This study concerns itself with material from the first set of administrations of the test, which took place over four days in November 1994 in test centres all around Australia. Approximately 7700 test-takers took part in these administrations.

1 Design of the step

The *step* includes writing, reading and listening components. This study considers only the writing component, which contains two tasks: 20 minutes for Task 1 (at least 100 words in length) and 25 minutes for Task 2 (at least 150 words in length).

The functional purpose of the two tasks was described as follows:

- Task 1: establishing and maintaining social relationships; giving/requesting information or explanations;
- Task 2: arguing or discussing an issue.

The rating scale used for the *step* was that developed for an earlier test used to assist in immigration decisions, the Australian Assessment of Communicative English Skills (*access*:) (Brindley and Wigglesworth, 1997). It contained four rating categories, each accompanied by descriptions at six levels, 0 to 5 (see Appendix 1), as follows:

- Task Fulfilment and Appropriacy (TFA);
- Conventions of Presentation (CoP);
- Cohesion and Organisation (C&O);
- Grammatical Control (GC).

Multiple forms of the test were prepared, to ensure test-takers had no prior knowledge of test content. All tasks were trialled before use in the test. Multi-faceted Rasch analysis employing the software FACETS (Linacre and Wright, 1992–96) was used for test analysis (for descriptions of the application of Rasch analysis in language performance assessment see McNamara, 1996; Weigle, 1998). Rasch analysis has the advantage of allowing estimation of the difficulty of each test task, and the harshness of each rater, and building these estimates into the calculation of the reported score for each test-taker. All test performances were rated by two raters. Rasch analysis also identifies test-takers whose scores do not conform to expected patterns, given the estimates of task difficulty and rater harshness. The scores for these test-takers are labelled as misfitting.

III Data collection

Four trained and experienced raters were selected for this study, from the entire pool of 65 accredited *step* raters. They shared similar backgrounds in terms of qualifications, teaching experience and experience as raters of ESL, as shown in Table 1. They had proved themselves consistently reliable during the two years of administrations of the *step* (Lumley, 2000). A group such as this should provide very good conditions for examining whether raters are capable of applying similar processes and criteria in their judgements.

Twenty-four scripts were selected[1] for this study, covering two forms of the test (out of the eight forms used in the November 1994 administrations). The tasks are given in full in Appendix 2. They were selected from scripts identified by the Rasch analysis as yielding misfitting (i.e., unexpected, or surprising) scores, because these were deemed to offer the greatest likelihood of eliciting comments from raters that would illuminate the rating process, including problems raters might encounter.

The first part of the data collection session was designed to simulate operational rating conditions, beginning with a reorientation similar to that conducted before operational rating. After this each rater rated a set of 12 scripts (two tasks each) as they would operationally. Each of the four raters then rated a second set of 12 scripts (two writing tasks each completed by the same 12 test candidates), but this time in addition provided concurrent think-aloud protocols describing the rating process, broadly following similar procedures to those proposed by Ericsson and Simon (1993). Instructions given to the raters

Table 1 Raters for this study: shared characteristics

- native speakers of English;
- aged 35–43;
- post-graduate ESL qualifications;
- ten or more years of ESL teaching experience;
- recent teaching experience with adults;
- experience teaching ESL overseas and in Australia;
- trained and experienced raters of other public assessments (e.g., IELTS, ACCESS, OET, ASLPR);
- trained and accredited as *step* raters at the same time;
- two or more years experience rating *step*.

Notes: IELTS = International English Language Testing System (Clapham and Alderson, 1996); ACCESS = Australian Assessment of Communicative English Skills (Brindley and Wigglesworth, 1997); OET = Occupational English Test (McNamara, 1990; 1996); ASLPR = Australian Second Language Proficiency Rating Scale (Ingram and Wylie, 1984).

[1]Only scripts from candidates who had passed the test were available for this study.

for this task are provided in Appendix 3. This study focuses on the data from stage 4 (Table 2), the think-aloud protocols produced as the raters awarded scores. The scores produced by the raters during Stage 4 of the data collection showed reasonable agreement among the raters, with (Pearson) correlations for scores between pairs of raters ranging between 0.71 and 0.91.

1 Data coding

It was necessary to develop a coding scheme that would adequately describe the raters' think-aloud (TA) data and address the research questions.

A broad orthographic transcription was carried out of each rater's talk, with transcription conventions indicating the source of raters' talk (test-taker script, rating scale or rater comments). The length of each rater's full set of TA protocols varied between approximately 10 350 and 17 500 words. Divisions of the TA protocols into text units for analysis were made according to the content of each unit. This is a pragmatic view, which recognizes that there is no single way to read a text and that division of the data into text units is ultimately an arbitrary act. The analyses conducted in this study attempt less to quantify instances of each behaviour, than to identify what range of behaviours can be observed, to consider whether or not raters demonstrate each behaviour, and to give a picture of the range of features that are examined by each rater.

Experimentation with coding schemes used by earlier researchers – including Huot (1988), Pressley and Afflerbach (1995) and Cumming (1990) – showed them to be unusable. This was consistent with the

Table 2 Summary of data

•	Four experienced, reliable raters;	
•	Two test forms (each with two tasks) = four different tasks.	
Stage		
1	Reorientation simulating operational conditions	Four practice scripts
2	Simple rating (no think-aloud)	scripts 1–12 × 2 tasks each = 24 texts
3	Practice think-aloud: rating	practice script
4	Rating plus think-aloud protocol	scripts 13–24 × 2 tasks each = 24 texts
5	Post-rating interview	

Note: Only data at Stage 4 is examined in this study.

view put by various researchers (e.g., Huot, 1988; Smagorinsky, 1994; Green, 1997; Torrance, 1998) that the categories needed to be developed to fit the data gathered in any context. A complex coding scheme was therefore developed to describe the think-aloud data from this study, consistent with the rating scale used in the *step*, and the particular focus in the research questions in this study. There was overlap with earlier schemes, especially those of Cumming (1990) and Cumming *et al.* (2001), but additional categories were identified, while others were perceived as less relevant. A preliminary survey of the data suggested that there were three very broad types of behaviours related to the rating process employed by raters:

- management behaviours;
- reading behaviours;
- rating behaviours: allocation of scores according to the four rating scale categories.

This distinction is consistent with the characterization of rating by Cumming *et al.* (2001: 15), in which they made:

> [a] basic distinction between *interpretation strategies* (or reading strategies, aimed at comprehending the composition) and *judgement strategies* (or evaluation strategies, for formulating a rating or score).

In total, 174 codes, grouped into six categories, were used. Four of these categories corresponded to the four categories of the *step* rating scale (see Appendix 4 for codes used relating to the assessment category, TFA), while the first category covered comments made during the first reading of the text, and the final category covered additional comments of various kinds. The coding scheme allowed analysis of the sequence of rating; the interpretations the raters made of each of the four analytic scoring categories in the *step* rating scale; and the difficulties raters faced during the rating process, together with the strategies they used to deal with these difficulties.

The author coded the entire set of texts. As a reliability check, a portion of the texts was coded by a second coder (an ESL teacher with postgraduate qualifications in applied linguistics). Of 300 coding decisions from five texts, agreement was 94%. Such a level of agreement suggested the coding scheme could be applied with adequate reliability.

IV The process of rating

This article now comments on raters' use and interpretations of the rating scale. First, similarities amongst the raters are noted, followed by their application of the scale.

1 Evidence of rater agreement

The study first aims to investigate to what extent the raters appear to agree on the rating criteria provided: this includes the extent to which they follow the given set of rating scale components, whether they comment on the same features of texts, and whether they stray outside the scale.

The first observation here is that most of the time the raters follow the rating categories provided, and in a very orderly way. There are some individual differences in the sequence with which they rate each category, but in simple terms the pattern follows three broad stages (Lumley, 2000), as shown in Table 3: first reading (or pre-scoring); scoring of the four categories in turn; and a final consideration of the scores given in Stage 2.

There is an overall shift during the rating of each text from attention on gaining an overall impression of the text during the first reading to a dual focus, on both text and scale, as raters allocated scores and reread sections of the text as necessary. At the end, there was sometimes consideration of the overall pattern of scores awarded. None of the scoring categories was neglected by any rater.

The view of the rating process which emerges here is consistent with the three-stage model described by Freedman and Calfee (1983), in which raters evaluate a 'text image' formed through reading the text itself, and filtered through their expectations, experience and background knowledge.

Table 3 Model of the stages in the rating sequence

Stage		Rater's focus	Observable behaviours
1.	First reading (pre-scoring)	• Overall impression of text: global and local features	• Identify script. • Read text. • Comment on salient features.
2.	Rate all four scoring categories in turn	• Scale and text	• Articulate and justify scores. • Refer to scale descriptors. • Reread text.
3.	Consider scores given	• Scale and text	• Confirm or revise existing scores.

Table 4 Components of scale for TFA

Rating category	Sub-category	Elaboration
Task Fulfilment and Appropriacy (TFA)	• relevance/appropriacy of content/ideas • meaning • vocabulary	• text relates to context • clarity/confusion/comprehensibility • presence of words • appropriate choice/errors/ effectiveness

2 *Scale content*

The scoring category, TFA, is discussed in detail in order to exemplify the ways in which the raters interact with the scale. The scale descriptors for TFA include clear references to various features of texts, as Table 4 shows.

The terms 'content' and 'meaning' are both relevant to this rating category. The distinction made here is between ideas and argument represented in the text (the content) and clarity, confusion or comprehensibility of what is said (the meaning). These scoring category components are clearly related to the scale descriptors, as can be seen in Table 5 (the relationship of the scale to the components identified in Table 4 is indicated in bold). Of the comments made by the four raters relating to the assessment of TFA, 704 out of 781 comments, or around 90%, fall into seven categories, as Table 6 shows. All of these comments clearly relate to the components of the scoring category, TFA, with the exception of the first code, which is a management statement, and the final code, which relates to the raters' reading process.

Table 5 *step* scale descriptors: TFA

0	No comprehensible English words. (Copied text should not be assessed.)
1	(a) **Text** is entirely **inappropriate** to given context or (b) predominantly **incomprehensible** although (c) a few **words or sentences may be present**.
2	(a) **Text relates** poorly **to** given **context** and is only **sporadically appropriate** or (b) **comprehensible**. (c) Some **appropriate vocabulary** within restricted **range**.
3	(a) **Text relates** in part **to** given **context** although (b) with some **confusion of meaning**. (c) **Appropriate vocabulary** used although there are considerable **errors**.
4	(a) **Text relates** generally **to** given **context** (b) with few **confusions of meaning**. (c) **Vocabulary** choices are generally **effective** although there are some **inappropriacies**.
5	(a) **Text relates** well to given **context**. It is thoroughly **appropriate** and (b) **easily understood**. (c) **Vocabulary choices** are **appropriate** and **effective**.

Table 6 Frequency of major codes for rater comments: TFA

Code	Comment focus	Total (percentages in brackets)
1	Scoring category nomination	46 (5.9)
2	Content/relevance only	255 (32.7)
3	Clarity of meaning only	95 (12.2)
4	Content/relevance plus clarity of meaning	37 (4.7)
5	Vocabulary	96 (12.3)
6	Overall category	148 (19.0)
7	Reading	27 (3.5)
Sub-total	These seven TFA codes	704 (90.1*)
Total	All TFA codes	781 (100.0)

Note: * Percentages do not add up exactly because of rounding.

The kinds of patterns that emerge in TFA are also seen in the other rating categories. The huge majority of comments relate to features explicitly contained in the scale (Table 7). Within the comments related to features explicitly referred to in the rating categories there were very few categories occurring with any frequency that were not shared by the raters as a whole: they all typically referred to the same sorts of things when discussing each rating category.

On one level, therefore, the rating procedure appears basically to work as intended by the test-developers, and we seem to have some basis for claiming that the contents of the scale are what raters attend to, and that the scale adequately describes the texts.

However, the fact that most comments refer directly to the scale does not mean that the rating process proceeds without difficulty. The following section considers a sample of the problems that may arise during rating. The examination of these problems shows that the issue of whether raters appear to understand and apply the rating scale descriptors in similar ways is much more complex than so far appears.

Table 7 Frequency of major codes during scoring

Rating category	Total comments	Comments explicitly related to scale features (percentages in brackets)
TFA	781	704 (90.1)
CoP	997	866 (86.9)
C&O	979	859 (87.7)
GC	990	906 (91.5)
Total	3747	3335 (89.0)

3 Conflict in scale wording

As we have seen, the category TFA explicitly includes both 'relevance of response' and 'clarity of meaning'. In the following extract of think-aloud data, Rater 3 (R3) describes the need to consider both of these components, but separately. For her, task fulfilment seems to be a necessary if not sufficient quality in a text, comparable to the 'hurdle' that test-takers must satisfy in the Foreign Service Institute (FSI) scale (Clark and Clifford, 1988).

R3–18A[2]

25. I'd better go over again to see if he's clear, if the meaning is clear,
26. that's the thing with the task fulfilment, **the first thing I look at is whether they've addressed the question,**
27. and then I have to, then I, then **that's, sort of seems to be separate from whether or not the meaning is clear,**
28. it seems, **you tend to end up dividing those two.**

In this next example, R3 shows how she resolves the conflict that may arise between these two features: in this instance she selects and justifies clarity of meaning as the overriding criterion, and gives the score, all in a few words. The frequent pauses shown here (each pause of 1 second is indicated by a '.') are suggestive of some hesitancy, but the rater manages the rating decision without much difficulty.

R3–13B

56. the, the, **the ideas are fine . worthy of a four . but . the meaning is not clear enough .**
57. and that's the thing that . affects task,
58. **task fulfilment has has to be . comprehensible meaning and that .. brings it down to a three again .**

For this rater, it appears, then, that either relevance or clarity may act as a hurdle. Vocabulary, the third component of TFA, although mentioned often enough in the data, appears to occupy a subsidiary role or to be subsumed under clarity of meaning. The emphasis on content and meaning exemplified above is very typical of what the raters say throughout the data.

Sometimes, however, the conflict that emerges between the two main scale features – relevance and clarity – is not so easily resolved. Several texts responding to the same task illustrate the variety of reactions and difficulties this conflict can cause. This task asks candidates

[2]Bold type is used in all quotations of rater talk to draw attention to the point under consideration; raters and scripts are identified by rater, script and task, as follows: R3–18A = Rater 3, Script 18, Task A; the figures in the first column identify the units of text (TU) into which each rating protocol is divided.

to enter a competition, for which the prize is a house. Their task is to 'describe the house [they] would like to win'. The main issue that arises is a mismatch between the raters' and the candidates' interpretations of what constitutes a relevant response to this task.

The problem is explained below, by Rater 1 (TU21–3), during her consideration of text 17A. We see that whereas the task rubric asks for a description, a number of the test-takers offer instead an argument for why they deserve or need to win the house. Here, vocabulary is not mentioned at all, while comprehensibility, although apparently perfectly acceptable (line 26), is ignored, and the script is given the lowest score of 1 for this category (line 33). The construct represented here now becomes something much narrower than the scale describes. Rather than falling back on her own internal criteria, the rater appears to be heavily bound by one particular aspect of the criteria stated in the scale descriptors: relevance. In fact, the issue of relevance was dealt with during training. Raters were provided with a set of guidelines to assist with various aspects of rating, including relevance, supplemented by raters' own notes taken during the training sessions, following examination of a range of examples of problematic texts. The test-developers' intention was to give guidance to raters on how they should treat relevance, and especially to encourage a broadly tolerant view amongst the raters of what should be seen as acceptable, while allowing them to penalize wholly or largely (apparently prelearnt) irrelevant answers. This purpose was discussed in some detail during training.

R1–17A[3]

20. well look,
21. **candidate is not answering the task,**
22. **the task is 'describe the house you would like to win',**
23. **this candidate is saying WHY he should win it**, so,
24. relates poorly,
25. but it is, sporadically appropriate or comprehensible
26. **well it's completely comprehensible**
27. but it is only mm, text is entirely inappropriate
28. well, no, **it's difficult here,**
29. **because the text can be entirely inappropriate but comprehensible, and this descriptor says**
30. entirely inappropriate to the given context OR predominantly incomprehensible
31. well it's . . . . . is it entirely inappropriate or relating poorly
32. and is only sporadically appropriate or comprehensible
33. **well I think actually it is entirely inappropriate,** [gives score of 1]

[3]Underlined text indicates a quotation or paraphrase from the rating scale.

34. the candidate has not answered the house he would like to win but merely why he should win a house . . .

The same problem arises with the next script. Again, the rater gives the lowest score point, although she seems to feel (for unstated reasons; see lines 18, 26) that neither time is the score very fair.

R1–18A

10. **no it's not answering the um, not answering the task at all, actually**
. . .

17. Okay, **I think the last time I a- I thin I, I said a one for this one,**
18. **seems rather harsh actually.**
19. text is entirely inappropriate to the given context
20. well look, it is entirely inappropriate,
21. it doesn't even relate poorly.
. . .

26. **well probably it should be two,**
27. **but um, really text is entirely inappropriate,**
28. **so I'll stick it at one**

She clearly places a major hurdle of relevance to her interpretation of the task in the pathway of the test-takers before they can score more than 1 or 2. Overall Rater 1 (R1) is more lenient than R3, but for this scoring category she applies a harsher criterion of relevance.

Rater 2 (R2) has a similar attitude to Rater 1, and awards a score of 2 for script 17A. She has more trouble in allocating a score when she considers the next script:

R2–18A

4. **this has a similar problem to the previous one,**
5. in that it doesn't describe the house you would like to win, it's trying to justify why I should win the prize,

Rater 2 then consults the raters' notes for guidance (lines 17–20 below), as well as her own notes taken during rater training and orientation sessions. As mentioned above, these notes – supplemented by examples and discussion during training – were intended to encourage a broadly tolerant view amongst the raters of what should be seen as acceptable (lines 18–20).

14. umm, now, it is comprehensible, **of course it's comprehensible,**
15. **but it's not relevant,**
16. and it's not like it's learned, it's not one of those 'springtime in Koreas'[4] . . .

[4]Text in quotation marks is extracted from raters' guidelines and supplementary notes.

17. 'two or more sentences are relevant'.
18. **'as long as text has something to do with topic that's okay'.**
19. 'if task asks them what they LIKE about the city and they write about the city that's okay ..'
20. 'as long as the text has something to do with the topic ..'

After commenting on the clarity of meaning in line 14 (above), she makes conflicting decisions (lines 21–23 below) about the relevance of the text, and notes the high level of subjectivity in interpreting the guidelines (line 22). Eventually she compares this text with the previous one (line 26–29) before allocating a score of 3, higher than for the previous candidate, apparently on the grounds of length of text.

21. **well, it's something to do with the topic,**
22. **this is a real value judgement here** I think
23. well, I don't think it does,
24. **I mean it does have something to do with it, but I don't think it has enough to do with it,**
25. **so I'm going to mark it down on that**
26. **so, what did I give that other guy, what did I give the other guy?**
27. **see, I only gave him a two,**
28. **and he hardly wrote anything at all,**
29. so I think we'll have to give it,
30. **I'm going to mark it down**
31. **because I don't think it's related**
32. text relates in part to given context although with some confusion of meaning,
33. **there's no confusion, but it's just not appropriate**
34. [inaudible]
35. **I don't like this criteria [said quietly]**
36. so I'm going to give it a three

This rater has a lot of trouble in allocating this score, and is dissatisfied with the rating scale (especially lines 22, 35). Her decision here is not clearly related to score descriptors, but rather to comparison with the previous test-taker's script: she appears to react in a holistic way to a feeling that this text is not worthy of a 4, but deserves better than the 2 she awarded to the previous script, which suffered from the same problem of questionable relevance. The wordings in the scale play little obvious part in the score decision. Likewise, the supplementary guidelines and training concerning relevance appear to have been largely set aside.

Raters 3 and 4 show different patterns again for texts 17A and 18A. R3 makes no comment on the relevance of the texts: her concerns were with clarity of meaning in both cases, while Rater 4 (R4) sees this as a problem only for the second of the texts, which he penalizes (giving scores of 2 for all categories). Clearly the raters differ in the way they apply the scoring category to these two scripts.

The raters' responses to these two texts (17A and 18A) can be summarized as follows:

- Raters 1 and 2 appeared to be concerned about the relevance of both texts, and commented on the similarities between them in this regard.
- R1 gave the lowest score (1) to both texts, on the grounds that they were similarly irrelevant. She seems to apply a 'hurdle' requirement to the tasks, unlike the other raters, although R3 does talk of this.
- R2 gave a higher score to the second one, apparently because she found it a better answer (certainly she commented that it was both comprehensible and longer than the earlier text). Both of her scores were more lenient than those of R1.
- R3 found no problem with the relevance of either text. Her concerns were with clarity of meaning in both cases, but she made no comparison between them.
- R4 apparently found no problem with relevance for the first task, but penalized the second one for lack of relevance, giving it a lower score. Again, he did not compare the texts.

It is clear from this variation in behaviour amongst the raters that there is actually a problem in the task construction, which did not emerge during trialling. However, the more significant point relating to rating is that the scale itself provides no guidance for this kind of interpretation, and the training and additional raters' guidelines, which were explicitly intended to address the issue of relevance, did not work as intended, and certainly were not interpreted consistently by the raters. When there is a problem they resort to other strategies, like heavily weighting one aspect of the criteria, or comparing scripts with earlier ones. The scoring decision appears not to be based on the scale. Such behaviours recur – in disparate and unpredictable ways – with all four tasks examined in this study, and with all four raters.

Because raters all react in different ways, both consistent measurement and consistent interpretation of scores given become very difficult. It might be argued that simply improving the scale wording would solve the problem. However, given the simplification and abstraction that scales necessarily entail, the conflicts that arise between different scale features are unlikely to disappear. Likewise, it might be suggested that training can eliminate difficulties such as this. We saw above, however, that explicit attention to the issue of relevance both in rater training and the notes given to raters failed to prevent the occurrence of these problems.

V The rating process revisited

In performance assessment, which relies on ratings, there is an assumption that if a rating scale is developed that describes writing texts in a valid way – and raters are adequately trained to understand its contents – then the scale will be used validly and reliably, and it will be possible to obtain good, or at least adequate, measurement. This suggests that raters must match the texts – which are produced in a myriad of ways, by a myriad of test-takers, from very different backgrounds, and in response to a range of tasks – to a single scale, which in a standardized test is usually written by somebody else. The assumption is that what the raters have to do is find the best fit: out of the scale descriptors presented to them, they have to decide which one best matches the text. Should the text be awarded a 3 or a 4? Raters should do this on the basis of common interpretations of the scale contents, developed as a result of training. What is observed here, when things go wrong, is a rather different process. In the preceding section raters' judgements appeared to be based on some complex and indefinable feeling about the text, rather than the scale content. The descriptions of the process of score allocation in fact allow us very little insight into the underlying processes raters go through. For example, raters referred to judgements they had already made about earlier texts as an additional basis for their score decisions. What we seem to observe is that the rater forms a uniquely complex impression independently of the scale wordings. Ultimately, however, they somehow managed in each case to *refer* to the scale content. This suggests that the role of the descriptors becomes one of articulating and justifying the scoring decision; in other words, the raters seemed to feel obliged to formulate the score decision in terms of the scale wordings, even when they experienced some level of discomfort with this. The next section illustrates further how raters use the scale as a tool for justification, as well as how the scale descriptors may be unhelpful when this process of justification is harder to manage.

1 Quantity of ideas

One significant feature missing from the *step* scale criteria – but which clearly forms part of the construct for raters – relates to the content of the writing, the quantity of ideas. This creates a number of significant difficulties for the raters, demonstrated in several of the texts. For example, Rater 3 encountered substantial difficulties with the following text (see Appendix 5 for the full script), and we see

once again how the rater's initial overall impression can cause a serious conflict with the scale descriptors, which the raters find hard to resolve. She states the basic conflict here:

R3–14A

25. um . so . this is a tricky one
26. because it's s- s' **it's** to- **very short.**
27. **but what . is written is actually quite . appropriate .**

She then tries to assess TFA, but finds it difficult, because of this conflict:

33. **so task fulfilment**
34. . . . er . it's not a . fulf- **it's not full enough . in ideas to be a four**
35. but it's . and it's, it's I'd say it's about a three
36. a two is not-.
37. because it, a two says it, text relates poorly,
38. well it definitely does relate . . .
39. it's a- four f- in a sense for task relating to the given context .. with few confusions of meaning
40. but **somehow or other**
41. **it's just not .. enough to be a four .**
42. . vocabulary choices are generally effective, although there are some inappropriacies
43. it's actually . **it's actually quite good .**
44. **I don't know why I wouldn't give it a four..**
45. I'll just keep it as . between three and four and I'll make, up my mind in a minute ..

So far, the text appears to match best level 4. But this doesn't satisfy her. Unable to make a judgement she moves to GC, and then C&O. Whereas the quantity of text or ideas seems clearly relevant to assessment of task fulfilment, it is less obviously part of cohesion and organization. This does not stop her looking for a reason to penalize the candidate under C&O, as we see here:

78. I'll probably give it . a three for cohesion and organisation
79. because **it just hasn't got enough** . it hasn't got ....
80. [whispered] oh why (not) why-y .... **why should I mark that one down to a three ..**
81. it's somewhere between a three and a four,
82. **it doesn't feel like it should be four but then when I look at it closely it's not ..**
83. **it's not too bad ....**
84. wide enough to park my c-truck
85. **I feel like I can't give it a four**
86. **because a four is something that should have fa- should have more in it.**

Her unease persists through the final two categories, CoP and GC:

92. I'll give it four for that. [CoP]

93. and I'll give it three-e
94. for .. grammar .
95. but it's actually quite good .....
96. **I feel like I have to kind of modify something,**
97. **I just can't, I can't quite give it . that much**
98. **I feel as if that's too high**
99. .....
100. oak in colour with very smooth surface . the colour should be Middle East product ...
101. **spelling,**
102. **I'll mark it . mark it down here because there's a . couple of spelling errors ..** [awards 3 for CoP]

This rater appears frustrated and really not helped by the scale. She appears to be seeking support for her intuitive view of what is right, and not getting it. For her, this is clearly not a matching activity at all. Instead, the scale is relegated to a subsidiary position, compared with her own view of what a level 4 should be. Her reason for penalizing the script ends up looking like a feeble pretext ('a couple of spelling errors', line 102). The final pattern of scores is 4–3–4–3, which seems to attempt to balance the good aspects of the script and its brevity. The rater in this extract appears very much to be influenced by a criterion that is not represented in the scale at all – length, or quantity of ideas – that pervades all the scores she gives.

Similar behaviour occurs with other texts, although R3 seems to find these less problematic.

2 Absence of explicit cohesive devices

A further area of comment concerns an additional feature absent from the scale, but which received many comments from the raters, namely, a lack of explicit cohesive devices. This is clearly relevant to the category of C&O, which refers to control of cohesive devices, but raters have to exercise their own interpretations in order to decide how their perceptions of this can be made to fit into the scale.

We see in this extract how, while scoring C&O, Rater 3 comments that the sentences are short (TU 56), that the ideas flow well (TU 58–59), but that the sentences should have been joined together more explicitly (TU 60). She then refers to the scale for a relevant description (TU 61).

R3–16B

56. his sentences are quite short.
57. I had to go,
58. the actual, from sentence, from idea to idea, from sentence to sentence is fine,
59. it flows.

60. although **he should have joined them, but he hasn't.**
61. **Absence of cohesive devices, where does that come into it?**
62 Inappropriate choices, oooh, generally cohesive, though problems may be noticed, organisation of ideas is mainly effective,
63. I think it's a four.

Here she has to decide whether this fits more likely with 'inappropriate choices' (3), or 'generally cohesive, though problems may be noticed' (4). She chooses the latter, knowing she has to align her impression with the scale descriptors; however, there is no real guidance here and it is an arbitrary decision.

R1 and R2 also comment on this issue, almost always while scoring C&O. Two more examples suggest this is a significant issue from the raters' point of view:

R1–14B

50. Cohesion and organisation.
51. Um, well, look, it's, there's not, **there's not a lot of, um, of overt cohesive devices,**

R2–22B

14. let's mark the coherence now
15. .. there's just, there's no.. **I can't see any conjunctions**
16. **this person doesn't like them**

What seems to be happening here is that the raters feel obliged to give explicit meaning and consideration to this category, C&O. They are required to articulate and justify their assessments under this category. However, much of the coherence of a text remains inexplicit (Widdowson, 1983). Since articulation and justification rely to a considerable extent upon explicitness, the raters seem to be forced into looking for explicit tokens, the most salient of which are conjunctive links. Consequently, raters also appear to use their absence as an associated criterion.

VI Conclusions

The preceding sections suggest that in response to the research questions, we may claim that although there appears to be some evidence that the raters understand the rating category contents similarly in general terms, there is also evidence that they sometimes apply the contents of the scale in quite different ways. They appear to differ in the emphases they give to the various components of the scale descriptors. Rather than offering descriptions of the texts, the role of the scale wordings seems to be more one of providing justifications on which the raters can hang their scoring decisions. In the context

studied here, at least, this seems to cast some doubt upon the idea that scales can assist us in understanding the constructs being measured by such ratings.

What implications may be drawn from this study? First, there are implications for our expectations of the training process. Rating is certainly possible without training, but in order to obtain reliable ratings, both training and reorientation are essential in order to allow raters to learn or (re)develop a sense of what the institutionally sanctioned interpretations are of task requirements and scale features, and how others relate personal impressions of text quality to the rating scale provided. Raters seem to be influenced, then, in the articulation of their ratings by a sense of the audience for the evaluation or rating, as Freedman and Calfee (1983: 93–94) noted. It seems clear that raters can be trained to use a scale, and to discuss the same sorts of features under specified rating categories. On the other hand, components of training sessions may go unheeded, partially heeded, or may take on proportions unintended by the trainer. It has been claimed in the past that the primary purpose of training is to forge common understandings, interpretations and agreement. This almost certainly happens, and the data from the current study seem to confirm this; however, it also seems that the primary purpose that scales and training end up serving is that of helping raters to articulate and justify their rating decisions in terms of what the institution requires, in the interests of reliable, orderly and categorized ratings.

Secondly, there are implications for understanding the role of the rater in performance assessment. The rater, not the scale, lies at the centre of the process. It is the rater who decides:

- which features of the scale to pay attention to;
- how to arbitrate between the inevitable conflicts in scale wordings; and
- how to justify her impression of the text in terms of the institutional requirements represented by the scale and rater training.

This study sheds more light on several earlier studies, and confirms their claims. For example, it appears clearer why Charney (1984) suggested that raters might apply 'idiosyncratic criteria', or why Vaughan (1991: 121) found that raters 'rely on their own styles of judging essays' when a text failed to match the scoring criteria. Cumming's (1990: 40) observation that expert raters spend much of their time 'classifying errors' perhaps relates to the requirement that they justify their ratings with observable features of the text, of which errors are the clearest evidence. Raters do not stop, as a result of training, having expert reactions, complex thoughts and conflicting feelings about texts as they read: we know that because they talk

about them in data such as in this study. However, they know that they have a particular job to do and, therefore, with the benefit of training, they just cope with this demanding task, shaping their natural impression to what they are required to do, in as conscientious a manner as possible, and using the scale to frame the descriptions of their judgements. We need experts to do this because the task is so complex: it does not require special training to read and form some sort of judgement of a text, but rating is considerably more complex than this.

Thirdly, there are implications for the validity of judgements and, particularly, the use of scales to describe test performance. It seems that scales are inevitably of somewhat limited validity, because of their inability to describe texts adequately. The role of a scale is rather as a tool for raters to use, to help in channelling the diverse set of reactions raters have when they read texts into narrower, more manageable, but by no means necessarily valid statements about them. Because this judgement is so complex, so multi-faceted, we can never really be sure which of the multitude of influences raters have relied on in making their judgements, or how they arbitrated between conflicting components of the scale. Likewise, we have no basis for evaluating the judgement that would have been made if a different scale were used.

I return to a comment from Huot (1993: 208). He expressed a fear that 'a personal stake in reading might be reduced to a set of negotiated principles, and then a true rating of writing quality could be sacrificed for a reliable one.' He succeeded in reassuring himself that this was not the case, because of the variability of reactions raters showed, and the range of topics they discussed. However, leaving aside the rather substantial question of whether or not there can ever be such a thing as a 'true' rating, he seems to have been closer to the mark than he realized, and ratings and scales represent exactly that: a set of negotiated principles that the raters use as a basis for reliable action, rather than a valid description of language performance.

Acknowledgements

This article is a substantially revised version of a paper presented at the 22nd Annual Language Testing Research Colloquium, Vancouver, Canada, March 2000. I would like to express my gratitude first of all to the four raters who worked with me on this project. I would also like to thank Tim McNamara, Liz Hamp-Lyons and Alan Davies, as well as three anonymous reviewers, for their helpful comments on earlier drafts of this article.

VII References

Alderson, J.C. 1991: Bands and scores. In Alderson, J.C. and North, B., editors, *Language testing in the 1990s*. London: Macmillan, 71–86.

Brindley, G. and **Wigglesworth, G.**, editors, 1997: *access: Issues in English language test design and delivery.* Sydney: (NCELTR), Macquarie University.

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

Clapham, C. and **Alderson, J.C.,** editors, 1996: *Constructing and trialling the IELTS test.* IELTS Research Report 3. Cambridge: The British Council, The University of Cambridge Local Examinations Syndicate and the International Development Program of Australian Universities and Colleges.

Clark, J.L.D. and **Clifford, R.T.** 1988: The FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status, and needed research. *Studies in Second Language Acquisition* 10, 129–147.

Connor-Linton, J. 1995: Looking behind the curtain: what do L2 composition ratings really mean? *TESOL Quarterly* 29: 762–765.

Cumming, A. 1990: Expertise in evaluating second language compositions. *Language Testing* 7, 31–51.

—— 1997: The testing of writing in a second language. In Clapham, C. and Corson, D., editors, *Encyclopedia of language and education, Volume 7: Language testing and assessment.* Dordrecht, Netherlands: Kluwer, 51–63.

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

DIEA 1995: 1 November 1993: Decisions, Fact Sheet 20, Revised 23 March 1995. Canberra: Department of Immigration and Ethnic Affairs.

Ericsson, K.A. and **Simon, H.A.** 1993: *Protocol analysis: verbal reports as data* (2nd edn). Cambridge, MA: MIT Press.

Freedman, S.W. and **Calfee, R.C.** 1983: Holistic assessment of writing: experimental design and cognitive theory. In Mosenthal, P., Tamor, L. and Walmsley, S., editors, *Research on writing: principles and methods.* New York: Longman, 75–98.

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

Hamp-Lyons, L. 1990: Second language writing: assessment issues. In Kroll, B., editor, *Second language writing*. Cambridge: Cambridge University Press, 69–87.

—— 1991: Scoring procedures. In Hamp-Lyons, L., editor, *Assessing*

second language writing in academic contexts. Norwood, NJ: Ablex, 241–76.

—— 1996: The challenge of second-language writing assessment. In White, E.M., Lutz, W.D. and Kamusikiri, S., editors, *Assessment of writing: politics, policies, practices*. New York: Modern Language Association of America, 226–40.

Hawthorne, L. 1996: The politicisation of English: the case of the *step* test and the Chinese students. *Australian Review of Applied Linguistics, Series S* 13, 13–32.

Huot, B.A. 1988: The validity of holistic scoring: a comparison of the talk-aloud protocols of expert and novice holistic raters. Unpublished PhD thesis. Indiana University of Pennsylvania.

—— 1990: Reliability, validity and holistic scoring: what we know and what we need to know. *College Composition and Communication* 41, 201–13.

—— 1993: The influence of holistic scoring procedures on reading and rating student essays. In Williamson, M. and Huot, B., editors, *Validating holistic scoring for writing assessment: theoretical and empirical foundations.* Cresskill, NJ: Hampton Press.

Ingram, D.E. and **Wylie, E.** 1984: *Australian second language proficiency ratings.* Canberra: Australian Department of Immigration and Ethnic Affairs.

Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V.F. and **Hughey, J.B.** 1981: *Testing ESL composition: a practical approach*. Rowley, MA: Newbury House.

Kroll, B. 1998: Assessing writing abilities. *Annual Review of Applied Linguistics* 18, 219–40.

Kroll, B. and **Reid, J.** 1994: Guidelines for designing writing prompts: clarifications, caveats and cautions. *Journal of Second Language Writing* 3, 231–55.

Linacre, J.M. and **Wright, B.** 1992–96: *FACETS.* Chicago, IL: MESA Press.

Lumley, T. 2000: The process of the assessment of writing performance: the rater's perspective. Unpublished PhD thesis, Department of Linguistics and Applied Linguistics, The University of Melbourne.

McIntyre, P.N. 1993: The importance and effectiveness of moderation training on the reliability of teachers' assessments of ESL writing samples. Unpublished MA thesis, Department of Applied Linguistics, University of Melbourne.

McNamara, T.F. 1990: Assessing the second language proficiency of health professionals. Unpublished PhD thesis, University of Melbourne.

—— 1996: *Measuring second language performance.* London: Longman.

Messick, S. 1989: Validity. In Linn, R.L., editor, *Educational measurement* (3rd edn). Washington, DC: The American Council on Education and the National Council on Measurement in Education, 13–103.

Milanovic, M., Saville, N. and **Shen, S.** 1996: A study of the decision-making behaviour of composition markers. In Milanovic, M. and Saville, N.,

editors, *Performance testing, cognition and assessment*. Selected Papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem. Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate, 92–114.

Moss, P.A. 1994: Can there be validity without reliability? *Educational Researcher* 23, 5–12.

—— 1996: Enlarging the dialogue in educational measurement: voices from interpretive research traditions. *Educational Researcher* 25, 20–28.

North, B. 1995: Scales of language proficiency. *Melbourne Papers in Language Testing* 4, 60–111.

North, B. and **Schneider, G.** 1998: Scaling descriptors for language proficiency scales. *Language Testing* 15, 217–62.

Pressley, M. and **Afflerbach, P.** 1995: *Verbal protocols of reading: the nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum.

Pula, J.J and **Huot, B.A.** 1993: A model of background influences on holistic raters. In Williamson, M.M. and Huot, B.A., editors, *Validating holistic scoring for writing assessment: theoretical and empirical foundations*. Cresskill, NJ: Hampton Press, 237–65.

Shohamy, E. 2001: *The power of tests: a critical perspective on the uses and consequences of language tests.* Harlow: Longman/Pearson.

Shohamy, E., Gordon, C. and **Kraemer, R.** 1992: The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal* 76, 27–33.

Smagorinsky, P. 1994: Potential problems and problematic potentials of using talk about writing as data about writing processes. In Smagorinsky, P., editor, *Speaking about writing: reflections on research methodology.* Thousand Oaks, CA: Sage, ix–xviii.

Spolsky, B. 1995. *Measured words.* Oxford: Oxford University Press.

Torrance, H. 1998: Learning from research in assessment: a response to writing assessment – raters' elaboration of the rating task. *Assessing Writing* 5, 31–37.

Vaughan, C. 1991: Holistic assessment: what goes on in the rater's mind? In Hamp-Lyons, L., editor, *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex, 111–25.

Weigle, S.C. 1994a: Effects of training on raters of English as a second language compositions: quantitative and qualitative approaches. Unpublished PhD dissertation, University of California, Los Angeles.

—— 1994b: Effects of training on raters of ESL compositions. *Language Testing* 11, 197–223.

—— 1998: Using FACETS to model rater training effects. *Language Testing* 15, 263–87.

Widdowson, H.G. 1983: *Learning purpose and language use*. Oxford: Oxford University Press.

Zhang, W. 1998: The rhetorical patterns found in Chinese EFL student writers' examination essays in English and the influence of these patterns on rater response. Unpublished PhD thesis, Hong Kong Polytechnic University.

Appendix 1 *Step* Writing performance criteria and descriptors

	0	1	2	3	4	5
Task fulfilment and appropriacy (TFA)	No comprehensible English words. (Copied text should not be assessed.)	Text is entirely inappropriate to given context or predominantly incomprehensible although a few words or sentences are present.	Text relates poorly to given context and is only sporadically appropriate or comprehensible. Some appropriate vocabulary within restricted range.	Text relates in part to given context although there are considerable errors. Appropriate vocabulary used although there are considerable errors.	Text relates generally to given context with a few confusions of meaning. Vocabulary choices are generally effective although there are some inappropriacies.	Text relates well to given context. It is thoroughly appropriate and easily understood. Vocabulary choices are appropriate and effective.
Conventions of presentation (CoP)	Absence of presentation conventions (spelling, punctuation, script, layout).	Very poor command of conventions of presentation (spelling, punctuation, script, layout) or sentences are present.	Uncertain command of conventions of presentation (spelling, punctuation, script, layout).	Adequate command of conventions of presentation, with some inconsistencies.	Generally good command of conventions of presentation although one area (spelling, punctuation, script or layout) may be weaker.	All aspects of presentation conventions (spelling, punctuation, script or layout) are handled skilfully.
Cohesion and organization (C&O)	Neither cohesion nor organization.	Very disjointed with minimal organization.	Limited control of simple cohesive devices, little awareness of appropriate organization of ideas relevant to this task.	Simple cohesion is controlled but problems of over use or inappropriate choices occur; there is some awareness of appropriate organization of ideas relevant to this task.	Generally cohesive, though some problems may be noticed in this area; organization of ideas is mainly effective.	Text is cohesive and organization is clear and appropriate to task.
Grammatical control (GC)	No grammar is evident.	Poor control of grammatical structures within this context.	Some control of grammatical structures suitable for this context but errors dominate.	Fair control of grammatical structures within this context but with considerable errors.	Generally good control of grammatical structures suitable for this context with a few obtrusive errors.	Competent control of grammatical structures appropriate to the context with only unobtrusive errors.

Appendix 2 *Step* tasks used for this study

Form C, Task 1

Situation:	You see a competition in the local newspaper to win a new house.
Picture:	drawing of a house, with the text: *'Win a new house. Describe the house you want to win and it could be yours.'*
Task:	Write to the newspaper editor. Describe the house you would like to win.

Form G, Task 1

Situation:	Last week you went on a day trip (for example, to the beach, zoo, mountains).
Task:	Write a letter to a friend. Say what you enjoyed most and why.

Form C, Task 2

Situation:	Your local public library is changing its closing time. Before, the library closed at 9.00 pm. Now, it will close at 6.00 pm. You want it to stay open until 9.00 pm.
Task:	Write a letter to the library. Say that you are unhappy about the change and give reasons (for example, work, family, study).

Form G, Task 2

Situation:	Your local council has $100 000 to build **either** i) a children's playground **or** ii) a car park. The council wants to know what people think.
Task:	Write a letter to the local council and say what you think. • Do you want the money spent on the children's playground or the car park? • Why?

Appendix 3 Instructions given to raters for think-aloud task

I am now going to ask you to rate a second set of 12 writing scripts. I would like you to rate them as far as possible in the usual way, that is, just as you have just rated the previous 12. However, there will be one important difference with this second batch: as I have previously mentioned, I am conducting a study of the processes used by raters when they rate writing scripts, and I would now like you to talk and think aloud as you rate these 12 scripts, while this tape recorder records what you say.

First, you should identify each script by the ID number at the top of the page, and each task within each script by number as you start to read and rate it. Then, as you rate each task, you should vocalise your thoughts, and explain why you give the scores you give.

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

Appendix 4 Coding scheme: sample

Codes used to categorize comments made during consideration of Task Fulfilment and Appropriacy (TFA)

Code	Category	Sub-category	Specific focus
1.1.0	TFA	TFA	nominates category
1.1.1	TFA	content	relevance/appropriacy/quality of argument/text – not scale-related
1.1.1a	TFA	content	relevance/appropriacy/quality of argument/text – scale-related
1.1.2	TFA	content	task requirement (reference to rubric)
1.1.4	TFA	content	reference to rater's instructions on relevance
1.1.5	TFA	content	personal reaction (interest, surprise, etc.)

Continued

Code	Category	Sub-category	Specific focus
1.1.6	TFA	content	summary of proposition(s)
1.1.7	TFA	content	quantity/length of text/ideas
1.1.7a	TFA	content	quantity/length of text/ideas (score-related)
1.1.8	TFA	contet? vocabulary?	appropriacy (unglossed)
1.1.9	TFA	content + meaning	relevance/appropriacy *plus* clarity
1.1.9a	TFA	content/meaning	relevance/appropriacy *plus* clarity (scale-related)
1.1.10	TFA	content/meaning	appropriacy vs. clarity – scale descriptor conflict
1.1.10a	TFA	content/meaning	appropriacy vs. clarity (scale-related) – scale descriptor conflict
1.2.1	TFA	meaning	clarity – not scale-related
1.2.1a	TFA	meaning	clarity – scale/score-related
1.2.2	TFA	meaning/script	illegibility/decoding or interpreting script (word level)
1.3.1	TFA	vocabulary	choice/range/accuracy/appropriacy) – not scale-related
1.3.1a	TFA	vocabulary	choice/range/accuracy/appropriacy) – scale/score-related
1.4.1	TFA	nonspecific comment/overall category (general)	overall quality/score – general comment (not scale-related)
1.4.1a	TFA	nonspecific comment/overall category (general)	overall quality/score – general comment (related to scale level and/or descriptor)
1.6.1	TFA	general	comparison with earlier text
1.6.1a	TFA	general	comparison with earlier text (score-related)
1.6.2	TFA	content	comparison with other test-takers in general
1.6.3	TFA	grammar	role of grammar in TFA score
1.6.4	TFA	CoP feature	CoP focus (as problem)
1.6.5	TFA	any	error classified in other rating category – example
1.6.6	TFA	task expectations	audience/register/formality/layout
1.6.7	TFA	cohesion	general comment
1.8.1	TFA	reading	rereading – part of text

Appendix 5 Full text of Script 14, Task 1 (Form C; see Appendix 2)

Script 14

Task A

street address
suburb
date

I am writing in response to your advertisement to describe a house I want to win.

The house should has at least five very big bedrooms with two story, two garages and a big swimming pool in the back yard. The house should be made of white brick with white iron fence around it. The drive away should be wide enough to park my truck. The stines should be made of good wood, oak in colour with very smooth surface. The carpet should be middle east product.

yours sincerely,

(90 words)