Assessment criteria for non-native speaker and native speaker essays: Do uniform standards work?

Sally O'Hagan
The University of Melbourne

Research on rater behaviour in the assessment of writing in English as a second language has largely focused on the concerns of language testing contexts; there has been comparatively little corresponding research into 'rater' responses in academic contexts. This article describes a small-scale survey of 21 lecturers—using their own assessment criteria—in the dental science, physiotherapy and education departments in an Australian university. The assessment of non-native speaker essays is discussed in relation to the question: in what ways are the assessment criteria for non-native speaker essays different from those for native speaker essays? While the findings show that, in general, the criteria are the same, they also reveal a tendency for lecturers to show leniency on some criteria for non-native speaker essays. Moreover, the findings point to a possible conflict between lecturers' actual responses to non-native speaker essays, and their feelings about how they *should* respond. The implications of such a conflict are considered in terms of the growing imperative—in the context of an 'internationalised' higher education environment—for lecturers to be able to engage confidently with a linguistically diverse student body. While this has relevance to higher education research and development, the findings also contribute to an under-researched area in the field of language testing: that is, what raters attend to when they make unguided assessments of writing.

1. Introduction

1.1 Background

Research on rater behaviour in the direct assessment of writing in English has explored the influence of writing features (such as content, conceptual organisation, mechanics) on rater judgment (Breland & Jones, 1984; Freedman, 1979; Raforth & Rubin, 1984) and has also examined the effect of rater variables (such as age, gender, professional background) on rater behaviour (Bridgeman & Carlson, 1983; Elder, 1992; Hamp-Lyons, 1991; O'Loughlin, 1992; Weigle, 1994). Raters' reactions to 'errors' (or the presence of non native-like

Melbourne Papers in Language Testing 1999 Volume 8.2 pp. 20-52. The Language Testing Research Centre, The University of Melbourne.

features) in writing by non-native speakers of English, have also been studied as a variable in rater response (Santos, 1988; Vann, Lorenz & Meyer, 1991). However, the behaviour of 'raters' working in academic settings, such as university lecturers, who respond to writing from a discipline-specific perspective, and moreover, without reference to a formal assessment tool, remains largely unexplored. With the continuing trend for universities in Australia (as in Canada, UK and USA) to recruit overseas students, coupled with the high rate of participation by migrant-origin students in tertiary education (see for example Hawthorne, 1997), university lecturers are now engaging with a highly culturally and linguistically diverse student body. This makes it an important time to seek a better understanding of how discipline (rather than language) specialists respond to writing by non-native speaking students.

Research on the experiences of non-native speaking overseas students has identified the kinds of academic problems that students face in a linguistically and culturally different setting. Ballard (1993) and Craswell (1992) are amongst those who discuss these issues in terms of how host institutions can implement ways to help students adjust to a new academic culture. However, as observed by Samuelowicz (1987), research has concentrated on the student experience, leaving little known about the experiences of the lecturers who teach these students. Samuelowicz's own survey of academic staff in an Australian university found perceptions of "language problems" in overseas students' written English, but indicates nothing about how the staff respond to these 'problems' (1987: 122). The aim of the research described here was to investigate responses to non-native speakers' academic writing of in terms of the assessment criteria used by lecturers to mark essays. The following questions were asked: i) Are the assessment criteria for non-native speaker essays different from those for native speaker essays? ii) If these criteria are not the same, in what ways are they different?

1.2 Researching rater behaviour: some methodological considerations

Using a questionnaire to survey professors in Canadian and USA universities about writing tasks assigned to students, Bridgeman and Carlson found that, when asked about any differences in the standards they use to evaluate the writing of native and non-native speakers of English, a "significant minority" of professors reported

that they judge some writing features more leniently for non-natives (1983: 30). The current study, with the use of a questionnaire, took a similar approach to investigating possible differences in lecturers' treatment of native and non-native speaker writing. However, from questionnaire data alone, there are limitations to what can be discerned about actual occurrences of contextualised behaviour. That is, respondents are constrained to reporting what they can perceive about themselves, which may not be the same as what they actually 'do' (Horowitz, 1986; Vann, Lorenz & Meyer, 1991), a problem which is illustrated in Breland and Jones' observation of a discrepancy between the "perceived and actual influence of essay characteristics" on raters (1984: 112). For this reason, the research design for the current study, combined questionnaire data with verbal protocols.

'Think aloud' protocol analysis, used widely in research on writing (for example, Flower & Hayes, 1980; Raimes, 1985) and reading (for example, Pressley & Afflerbach, 1995; Wyatt et. al., 1993) is also proving to be an important tool in language testing research for studying rater behaviour in the assessment of writing (for example, Huot, 1990; Vaughan, 1991), and more recently, speaking (for example, Brown, 2000; Brown et. al., forthcoming; Meiron, 1998). Whilst not without their own limitations, relating primarily to accuracy of data, and to comprehensiveness as records of behaviour (an issue which will be returned to in section 4.2), verbal protocols are potentially rich sources of information about what raters attend to, what they think about, and what they do during the rating process, rather than what they remember of it at some other time (such as when responding to a questionnaire). The issue of accuracy relates to the potential for 'think aloud' procedures to slow, or alter task performance, and has long been discussed in the literature as a problem (see Nisbett & Wilson, 1977). However, in more recent debate, an understanding of the effect of task type has emerged, bringing many to the view that by using tasks that are routine or familiar (Pressley & Afflerbach, 1995) and language-based or well described in language (Johnson & Briggs, 1994; Brown, 1987; Geisler, 1994) the impact of 'think aloud' procedures on task performance is minimised. Certainly, 'think aloud' techniques have been used to gain considerable insights into rater decision-making processes by the likes of Cumming (1990), Milanovic, Saville and Shuhong (1996) and Weigle (1994), who investigated ESL/EFL composition raters' use of the scoring instruments provided to them. In writing assessment and other areas of testing, such studies provide opportunities for

understanding rater contribution to measurement 'error' in relation to test validity: are raters using the scoring instrument 'properly', and aside from specified criteria, what else is informing their judgments?

In 'real world' academic settings, like that of the current study, it is the second of these concerns that is of most interest. University lecturers are experts in their field, but they are not trained as composition raters (nor are they subject to the same kind of expectations of reliability) – in general, lecturers do not mark essays with reference to formally articulated assessment criteria; if such criteria are available, lecturers either are not obliged to use them, or may use them in combination with unspecified criteria of their own. Although the extent to which lecturers meet with colleagues to moderate or discuss their judgments varies within and between institutions, essay marking of the kind investigated here, is typically an individual affair.

2. Method

2.1 Overview of design

The research design utilised two stages of data collection to survey university lecturers: administration of a questionnaire to all subjects (Stage 1); and collection of concurrent verbal reports of non-native speaker essay marking protocols from a subset of lecturers in the sample (Stage 2). In Stage 1, written questionnaires for self-completion were distributed to lecturers who were both eligible and available (see section 2.2, below) to take part in the study: 10 in physiotherapy, 12 in dental science and four in education. Of these, responses were received from nine lecturers in physiotherapy, nine in dental science, and three in education. Of the 21 respondents, 12 expressed interest in taking part in the second (verbal protocol) stage of the study. Three lecturers—the first available from each department—were selected to participate in Stage 2.

2.2 Subjects

The subjects were lecturers in three academic departments: physiotherapy, dental science, and education. The two health science departments were targeted on the basis of a perceived need and interest amongst their staff for research on teaching non-native speaker students, and because of the high participation rate of non-

native speaker students in these two departments. The following selection criteria applied to individual participants within each department: in courses taught by participating lecturers, a significant proportion of the students were non-native speakers; and formal assessment of student performance included essays of at least 500 words. Lecturers in the third discipline—education—were selected for the study using the above criteria. Inclusion of this group was intended to enable comparison of responses of language- and non language-trained specialists – although participation rates in education were low, at least two of the lecturers in education were experienced in the areas of literacy and language education.

2.3 Materials and Procedures

2.3.1 Stage 1: Questionnaire

The questionnaire was developed after conducting exploratory interviews with lecturers across 10 academic disciplines in departments not involved in the main study. In these interviews, staff were asked to describe their general approach to essay marking plus their treatment of essays which, in their view, show distinctive language problems, or non native-like features. These descriptions informed the design and content of the survey instrument used in Stage 1, and shown at Appendix A. Items consisted of a mixture of open and closed questions, and were informed in a general sense by Bell's (1993) guidelines on questionnaire design. Respondents were asked to describe their own marking procedures and assessment criteria, and to compare their responses to writing by non-native and native-speaking students. They were also given a list of writing features and—reporting separately on their treatment of non-native and native speaker essays—asked to use a rating scale to show the importance of each as a criterion in their own marking scheme. The list of features was derived from previous research on the influences of content, organisation and grammar on rater judgment (including Breland & Jones, 1984; Bridgeman & Carlson, 1983; Freedman, 1979; Huot, 1990a). Specifically, the categories of writing features used in Bridgeman and Carlson's survey (see 1.2, above)— which had been trialed and tested to make sure it was "free of linguistic jargon" and therefore comprehensible to respondents (1983: 12)—were chosen as a model which was modified (on the basis of the assessment criteria most widely cited in the exploratory interviews) to ensure that conventions and terminology were appropriate to the local setting.

2.3.2 *Stage 2: Verbal Protocols*

In Stage 2, audio-recordings were made of participants verbalising their thoughts aloud whilst marking non-native speaker essays. The task for participants was to read and award a mark/grade to as many essays as they would in a normal marking session without taking a break, determined to be a period of 30–45 minutes. The number of essays marked during this period was dependent on the length of the essays, and on the speed and style of each participant. As all had substantial experience marking essays in their field of expertise, the task given to participants was routine. The essays were all authentic texts that had been written by students in the relevant departments for course assessment purposes, and consisted of discursive essays of 1,000 to 3,000 words on a specified topic. Each participant supplied their own 'pool' of essays to draw on during the session. These had been written by students in the same course and on the same topic. Participants had not previously read the essays, but were familiar with the topic and requirements because the essay had been set for their own course or a course they had taught before. From this 'pool', participants selected essays that, according to their knowledge or judgment, had been written by a non-native speaker.

The instructions given to participants (shown at Appendix B) were based closely on those used by Perkins (1981: 33) and Geisler (1994: 260–261) and were supplemented with ideas drawn from Berkenkotter (1981). To maximise reporting about the specific marking session at hand, rather than generalisations from many marking sessions, participants were provided with a list of suggested talking points (also shown at Appendix B) in addition to the general instructions, a procedure also followed by Olson, Duffy and Mack (1984). This list of written prompts was theoretically motivated to help elicit information relevant to the research questions. It also served the purpose of facilitating continuous talk throughout the marking session – as the researcher was not present, no spoken prompts were given to keep participants talking throughout the session. Participants were instructed to address the points on the list only when or if they felt they needed something to prompt their thoughts.

2.4 Data Analysis

2.4.1 Questionnaire data

Simple statistical procedures were performed on lecturers' ratings of writing features (indicating the importance of each feature as an assessment criterion). Descriptive statistics were used to establish the order of importance of writing features overall, and to compare this order across the three departments. To compare the importance of these features when used to evaluate non-native and native speaker essays, each respondent's rating of each criterion for non-native speaker essays (expressed as a score from 1 to 5) was subtracted from the corresponding rating for native speaker essays, thus generating a *difference score* which is zero, negative or positive. For example, a rating of '3' for native speaker essays, compared with '3' for non-native speaker essays, is expressed as *no difference* (d = 0). Similarly, a rating of '1' (very important) for native speaker essays, compared with '4' (not very important) for non-native speaker essays is given a negative score which is expressed as *difference* (d<0). Using this formula, an instance of a score of *difference* (d<0) shows that a criterion has been rated as less important for non-native speaker essays (or as more important for native speaker essays). Of course, it is possible for the scores to go in the opposite direction: for example, a rating of '4' (not very important) for native speaker essays, compared with '1' (very important) for non-native speaker essays would generate a positive score, or *d>0*. Each type of difference score (d=0, d<0, and d>0) was tallied for each feature.

2.4.2 Verbal Protocols

Broad transcriptions of the protocols were completed prior to coding and analysis. Interpretative analysis of the protocols was carried out subsequent to encoding, according to procedures suggested by Berkenkotter (1983), Geisler (1994) and Swarts, Flower and Hayes (1984). As pointed out by Ericsson and Simon, whilst 'think aloud' research is characteristically data-driven, the methodology nevertheless will "rest on a set of assumptions" about the object of study (1993: 263). For Wyatt et. al. (1993) for instance, these assumptions concern the nature of reading behaviours. In the current study—while encoding was a primarily data-driven activity—the subsequent analysis was sensitive to themes identified in both the literature discussed above, and the data from Stage 1.

Strauss and Corbin's (1990) methodology of procedures for open coding of qualitative data (a grounded theory approach) provided a framework for developing the coding scheme. A similar approach was followed by Milanovic, Saville and Shuhong (1996), and is recommended by Green (1998) in her handbook of verbal protocol methods. Firstly, each protocol was divided into codable segments using the criterion that each segment would consist of discourse about a single topic. This procedure produced segments consisting of single or (commonly) several utterances, often bounded by long (timed) pauses and/or by topic shift markers such as 'right', 'so', 'again', 'and yet'. Next, the segments were organised into like groups, or coding categories, which were then progressively modified with repeated examinations of each protocol until all segments could be accounted for by the final coding scheme (shown at Appendix C). Then, a procedure resembling the "pattern coding" described by Miles and Huberman (1994: 69) was followed. That is, where appropriate, the coding categories were organised into groups consistent, or matched, with analytical categories appropriate for the questionnaire data. This process of 'matching' was employed to enable a consistent analytical focus across the two stages of this study, thus allowing a comparison of the behaviour recorded in the protocols with lecturers' generalised descriptions of their behaviour (as provided in their responses to questionnaire items). Finally, a subset of the data consisting of 189 segments out of a total of 375 (or 50.40% of all segments), was coded by an independent coder. The level of inter-coder agreement for the data subset was 82%.

The data corpus of 3 protocols (one from each participant) consisted of verbal reports of the marking of a total of 11 different essays: 5 dental science, 4 physiotherapy, and 2 education essays. Encoding yielded data in the following quantities for each protocol: dental science – 142 segments; physiotherapy – 122 segments; education – 152 segments. To some extent, the number of segments in each protocol is simply a reflection of individual differences between participants: how much each had to say, and how concisely or otherwise they said it. Note that while the lecturer in education marked fewer essays than the lecturers dental science and physiotherapy, the total number of data segments in the education protocols is the highest. Apart from the individual differences just mentioned, it is also reasonable to expect longer protocols about longer essays – the education essays were 3,000 words compared with 1,500 or less in the other two disciplines. Another factor which

differentiates the education protocols in this study, is that the essays were written by Masters level students, whereas the essays in the other two disciplines were all written by undergraduate students. It is not unlikely that greater complexity in the Masters essays provided the lecturer with a greater stimulus for her verbal reports. Note that, of the data groups shown in the code list (Appendix C), this article focuses only on the first two: analytic evaluations, and reading strategies (a full report appears in the form of an unpublished dissertation).

3. Results

3.1 Criteria used to mark essays

Content and structure/organisation are the writing features most frequently and most widely reported as assessment criteria: Table 1 shows the features that were cited by lecturers (and in what frequencies) when they were asked to specify the assessment criteria they use to mark essays; Table 2 shows which features were evaluated in the verbal protocols, how frequently, and how these evaluations were distributed across the total number of essays marked.

Table 1: Features cited as lecturers' own assessment criteria

	All Departments ($N = 21$)	Dental Science ($n = 9$)	Physiotherapy ($n = 9$)	Education ($n = 3$)
content	20	9	8	3
structure/organisation	11	3	6	2
presentation	9	4	2	3
research	8	2	3	3
written communication skills	6	2	3	1
referencing	4	1	2	1
grammar	3	-	3	-
spelling	2	-	2	-
punctuation	1	-	-	1

Table 2: Occurrences of analytic evaluations of writing features in verbal protocols

	All Departments	Dental Science	Physiotherapy	Education
content	89 (11)	42 (5)	22 (4)	25 (2)
structure/organisation	42 (10)	14 (4)	20 (4)	8 (2)
grammar/sentence structure	25 (9)	3 (3)	11 (4)	11 (2)
overall writing ability	15 (6)	-	11 (4)	4 (2)
referencing	13 (3)	-	2 (1)	11 (2)
vocabulary	6 (3)	-	1 (1)	5 (2)
length	3 (3)	-	2 (2)	1 (1)
legibility	2 (1)	2 (1)	-	-
register	1 (1)	-	-	1 (1)
	N = 196 (11)	n = 61 (5)	n = 69 (4)	n = 64 (2)

N = frequency of evaluation; (N) = number of essays in which evaluation occurs at least once

As shown in Table 1, content, structure/organisation and presentation were the most commonly cited features overall (i.e. across all three departments). Research, overall written communication skills, and referencing were the next most commonly cited criteria, followed by grammar, spelling and punctuation. As shown in Table 2, the features most widely evaluated in the protocols were content, structure/organisation, grammar/sentence structure and overall writing ability, all occurring in more than half of the essays and in the highest frequencies. Referencing, vocabulary and length follow as the next most widely evaluated features, all evaluated in three essays and in lower overall frequencies than the first group. Finally, references to legibility and register were made least of all: in only one essay each, and in the lowest overall frequencies.

3.2 Relative importance of writing features as assessment criteria

Overall, content, structure/organisation and style were rated and ranked by lecturers as the features that are most important to them as assessment criteria for native speaker and non-native speaker essays. Table 3 is derived from lecturers' ratings of each feature on a five point scale (from 'very important' to 'not at all important') to indicate how important each feature is to them as an assessment criterion. The descriptive statistics of these ratings (see Appendix D) were used to determine the position of each feature on the 'most to least important' hierarchy shown in Table 3. As well as rating each on a five point

scale, lecturers were asked to choose their 'top three' most important features. Table 4 shows the frequency with which lecturers ranked each feature as their first, second and third most important assessment criterion.

Table 3: Relative importance of features used as assessment criteria

	Most important criterion →							Least important criterion
All Depts.	NS	content	struct/org	style	gramm/ss	spelling	vocab	punct
(*N* = 21)	*NNS*	*content*	*struct/org*	*style*	*vocab*	*gramm/ss*	*spelling*	*punct*
Dental Sc.	NS	content	struct/org	style	vocab	gramm/ss	spelling	punct
(*n* = 9)	*NNS*	*content*	*struct/org*	*style*	*vocab*	*gramm/ss*	*spelling*	*punct*
Physioth.	NS	content	struct/org	style	gramm/ss	spelling	vocab	punct
(*n* = 9)	*NNS*	*content*	*struct/org*	*style*	*gramm/ss*	*vocabulary*	*spelling*	*punct*
Education	NS	content	struct/org	gramm/ss	style	spelling	punct	vocab
(*n* = 3)	*NNS*	*content*	*struct/org*	*gramm/ss*	*style*	*punct*	*spelling*	*vocab*

struct/org – structure/organisation; gramm/ss – grammar/sentence structure;
vocab – vocabulary; punct – punctuation
NS – native speaker essays; NNS – non-native speaker essays

Table 4: Features ranked as the three most important assessment criteria

		content		struct/org		style		gramm/ss		vocab		punct	
	Rank	NS	NNS	NS	NNS	NS	NNS	NS	NNS	NS	NNS	NS	NNS
All Depts.	1st	17	18	4	3								
(*N* = 21)	2nd	3	2	12	13	3	4	1	-	1	1		
	3rd	1	1	4	4	7	7	5	6	1	1	1	-
Dental Sc.	1st	7	7	2	2								
(*n* = 9)	2nd	1	1	4	5	2	2	1	-	1	1		
	3rd	1	1	2	2	3	3	-	1	1	1	1	-
Physioth.	1st	7	8	2	1								
(*n* = 9)	2nd	2	1	6	6	1	2						
	3rd			1	1	4	4	3	3				
Education	1st	3	3										
(*n* = 3)	2nd			2	2								
	3rd			1	1			2	2				

Additional features ranked '3rd', all with frequencies of '1' for both NS and NNS - "legibility" (dental science);
"interesting to read" (physiotherapy); "use of own voice" (education).
struct/org – structure/organisation; gramm/ss – grammar/sentence structure
vocab – vocabulary; punct – punctuation
NS – native speaker essays; NNS – non-native speaker essays

According to an impressionistic overview of Tables 3 and 4, content, structure/organisation and style are the most important criteria; grammar/sentence structure and vocabulary follow as the next most important, while spelling and punctuation figure as the least important.

3.3 Differences between lecturers' treatment of non-native and native speaker essays

3.3.1 *Differences in the importance of the writing features used as assessment criteria*

In Table 3 (above) some differences can be observed between the order of features on the hierarchy of assessment criteria for native speaker essays and the parallel hierarchy for non-native speaker essays i.e. towards the 'least important', or lower, end of these hierarchies. Through analysis of the raw scores from which the hierarchies were derived, the differences between the hierarchies can be quantified in terms of the difference scores described in 2.4.1, above. The frequencies of each type of difference score for each feature are shown in Table 5.

Table 5: Differences between importance of features as assessment criteria for non-native speaker compared with native speaker essays

		content	struct/org	style	vocab	punct	grammar/ss	spelling
All Depts.	difference	-	2	5	5	9	11	11
(*N* = 21)	no difference	21	18	15	15	10	10	9
	d>0	-	1	1	1	2	-	1
Dental Sc.	difference	0	1	3	1	3	3	3
(*n* = 9)	no difference	9	7	6	8	6	6	6
	d>0	0	1	0	0	0	0	0
Physioth.	difference	0	1	1	4	5	6	7
(*n* = 9)	no difference	9	8	7	5	4	3	2
	d>0	0	0	1	0	0	0	0
Education	difference	0	0	1	0	1	2	1
(*n* = 3)	no difference	3	3	2	2	0	1	1
	d>0	0	0	0	1	2	0	1

struct/org – structure/organisation; grammar/ss – grammar/sentence structure
voab – vocabulary; punct – punctuation
difference – d<0; no difference – d=0

The high frequency of scores of *no difference* for both content and structure/organisation, together with a correspondingly low

frequency of scores of *difference* indicates that there is little or no difference between the importance of these features for non-native and native speaker essays. Scores of *difference* for both style and vocabulary show that these features are less important for non-native speaker essays (or more important for essays by native speakers) for five of the 21 lecturers. Punctuation, grammar/sentence structure and spelling have the lowest frequency of scores of *no difference*, together with the highest frequencies of scores of *difference*. Thus, the scores in Table 5 suggest an overall trend for the more 'surface level' features (punctuation, grammar/sentence structure and spelling) to be given less weight for non-native than for native speaker essays, with content and structure/organisation tending to be just as important as assessment criteria for both types of essays.

Note that small number of positive scores ($d>0$) appears in Table 5. At face value (according to the analytic framework used in this study), these scores would suggest that for some lecturers, certain features are more important for non-native than for native speaker essays. However, not only are these scores both few in number and highly anomalous with the trend described above, this interpretation is problematic (or of limited meaningfulness) in the context of the study as a whole i.e. what does it mean for a feature to be 'more important' to a lecturer when they mark non-native speaker essays? It is unlikely that 'more important' means a higher standard is expected in non-native speaker essays. Another possibility is that it is 'more important' for the non-native speaking students (to try harder than native speakers). Alternatively, 'more important' could mean that problems are more likely to occur in non-native speaker essays, leading lecturers to be more aware of certain features. This problem with understanding what these scores 'mean' may indicate that better clarification of the terminology in the questionnaire was required ie. did these scores arise from respondent 'error'/failure to understand the question? (albeit the risk that researchers and respondents will not 'speak the same language' can never be removed entirely). All the same, that prior validation of the original questionnaire item by Bridgeman and Carlson (1983), plus piloting of the version adapted for this study seems to have been a reasonable precaution, is borne out in the fact that the interpretation of 'important' by the majority of lecturers was consistent overall. Even amongst those whose ratings produced $d>0$ scores, the interpretation of 'important' was generally consistent with that of the rest of the sample. For example, one respondent who rated style as more important for non-native speaker essays, at the same time rated grammar/sentence structure, spelling

and punctuation as less important, and reported being lenient for grammar and spelling, thus suggesting that 'importance' is a measure of stringency or how heavily a criterion is weighted in this lecturer's marking scheme.

3.3.2 *Use of leniency on some assessment criteria*

In other differences, leniency on certain criteria is the most commonly reported difference between lecturers' treatment of non-native and native speaker essays, reported by 13 out of 21 lecturers in Stage 1. Examples follow in Table 6, below. Leniency was most commonly reported for sentence level features such as grammar/sentence structure, spelling and vocabulary (by nine out of twenty-one). In most cases (six), these reports included the proviso that the same quality in content is expected, or that the message must still be communicated clearly (as in the examples given in Table 6). Fewer lecturers (four) reported using leniency for discourse features, such as overall structure, coherence, development of argument and cohesion.

Table 6: Lecturers' reports of leniency on assessment criteria for non-native speaker essays

Feature/level	ID	Example (Stage 1)
sentence level language use	28	I apply the same content criteria, but tend to make allowances for less experience with the English language on the formal criteria.
	1	Sometimes spelling and grammar are incorrect – therefore I am generally more lenient if I feel I can get the gist of what the student is trying to say.
discourse features	12	More accepting of a less elaborated argument or shorter piece of work.
	13	Occasionally I may request interview with [non-native speaker students] to check on cohesiveness of argument.

ID – lecturer identification code

The exercise of turning to the verbal protocols for instances of lecturers talking about being lenient/stringent on various criteria for non-native speaker essays (or examples of the behaviour described by lecturers in Stage 1), was particularly illuminating in what it revealed about how leniency is articulated through certain strategies adopted by lecturers in responding to non-native speaker essays. In the examples in Table 7 (below) particular reading strategies are used as a way of dealing with language problems. (In the examples from lecturers 4 and 15, data is drawn from both Stages 1 and 2). In the first example (lecturer 4), leniency on grammar, seems to 'translate' in

practice into taking more 'care' in reading the essay. On other hand, in the second example (lecturer 15), the response is to 'go quickly' in search of the main points. A greater reliance on finding 'key points' for evaluating non-native speaker essays is a strategy described by others in Stage 1 – in the remaining 2 examples in Table 7, there is a sense that this strategy of 'looking harder' for the content of an essay, is about giving non-native speakers a fair opportunity. While there were no reports of leniency for content, in the examples in Tables 6 and 7 lecturers are in a sense privileging content over other features in non-native speaker essays, either by showing leniency for another feature which is weak, or by being more attentive to content as a way of dealing with the difficulties of reading a text with grammatical or organisational problems.

Yet the use of these strategies is not always unproblematic, as can be seen from the examples in Table 8, below. For instance, the comments by lecturer 15 (Stages 1 and 2), who reported that he is more apt to look for key words in non-native speaker essays, express not only ambivalence about the success of this strategy, but also concern about whether it is equitable: in the questionnaire, this lecturer expressed reservations about relying on key words; he then went on to express doubts in his verbal report about his own decision (on finding the essay difficult to understand) to skim read for key words.

Table 7: Strategies for responding to non-native speaker essays

ID	Data source	Example
4	Q'nnaire (Stage 1)	Slightly more lenient, mainly in grammar; I look more message rather than just concentrating on simple grammatical mistakes.
	Protocol (Stage 2)	Although there are a lot of grammatical errors, the section on physiotherapy treatment is better factually, so I'm starting to forgive the grammatical errors, and read a bit more carefully.
15	Q'nnaire (Stage 1)	[Non-native speaker] students tend to be assessed more on key word use.
	Protocol (Stage 2)	The grammar makes the sentences more difficult to understand, so when I see a paper like this I tend to just go very quickly and look for the key points.
22	Q'nnaire (Stage 1)	I look particularly hard for key words and may re-read the answer a number of times to ensure every opportunity possible for the student.
8	Q'nnaire (Stage 1)	I try hard to mark on content. I also know the students well and I try to unravel their answers with that background.

ID – lecturer identification code

Table 8: Lecturers' comments on the difficulties of assessing non-native speaker essays

ID	Data source	Example
15	Q'nnaire (Stage 1)	Key words alone obviously do not indicate understanding. It is difficult to assess the discussion.
	Protocol (Stage 2)	[To go very quickly and look for the key points] may advantage them in a way because I'm not too concerned with how they've put it down.
2	Q'nnaire (Stage 1)	Sometimes difficult to understand what is written – may appear that the student has knowledge, but displays difficulty expressing it.
4	Protocol (Stage 2)	This person might have a better understanding than the way he writes.

ID – lecturer identification code

The examples shown in Table 8 are indicative of doubts about whether it's fair, or appropriate, to try to read through grammatical errors to uncover what it is that the student 'really' means. A sense of unease can be detected in the comments: although there were no explicit reports of 'leniency on content', some lecturers decide to make allowances for poor written communication skills, and at the same time, experience doubts about whether they are able to judge what the student really does and doesn't know.

3.4 Differences between departments

Priority has been given throughout this section to reporting the overall findings ie. across all three departments. A brief mention of some of the departmental differences that were observed is offered here as speculation, since the small *n* sizes in this study provide no basis for generalisation. In any case, compared with the findings of previous research, the observed differences are inconclusive (readers may wish to review Tables 1–5). The most salient differences are in the education department data where comparatively more importance is placed on grammar/sentence structure, and comparatively less importance on vocabulary than in either of the health science departments (refer to Table 3). Bridgeman and Carlson (1983) postulate from their findings that people teaching in language focused disciplines are those most likely to attach greater importance to the mechanics of writing. On the other hand, Santos (1988) and Vann, Lorenz and Meyer (1991) found that raters in the humanities/education/social sciences were more likely to be tolerant

of mechanical errors in ESL writing than their counterparts in the physical/biological sciences. Regarding the difference in the importance of vocabulary, one might speculate that the basis for this could lie in a comparatively greater importance of specialist terminology in dental science and physiotherapy: the tendency that was found in this study for a greater reliance on key word use for evaluating content (described in section 3.3.2, above) occurred in the health sciences only, suggesting that further investigation would be warranted. In future research, a larger sample, stratified to allow for interdisciplinary comparison, would make this possible.

4. Discussion

4.1 Summary and discussion of findings

With respect to the research questions posed at the end of section 1.1 (above), the key findings are:

- In general, the assessment criteria applied to native speaker and non-native speaker essays are the same, but some writing features are considered less important as assessment criteria for non-native speaker essays. In particular, there is a tendency for punctuation, grammar and spelling to be considered less important for non-native speaker essays. The same is true for vocabulary and style, but the tendency is less marked.
- While the criteria themselves are not different, marking schemes are commonly modified in some way for non-native speaker essays. In particular, there is a tendency to show leniency on some assessment criteria for non-native speaker essays, commonly for grammar, spelling and vocabulary, and occasionally for structure/organisation.

In relation to other features then, less evidence was found for differences between the use of content and structure/organisation as assessment criteria for non-native speaker essays, compared with native speaker essays. With ample precedent in the research literature that content and structure/organisation are the features that have the most influence on rater judgment (for example, Breland & Jones, 1984; Bridgeman & Carlson, 1983; Freedman, 1979; Huot, 1990a), it is not surprising that, for the lecturers in this study, content and structure/organisation are the most important criteria (in terms of how commonly they are used as assessment criteria, and how

important they are considered to be compared with other criteria) used to make judgments about the quality of all essays, regardless of the writer's language background.

What is interesting however, is the issues arising from the differences that do occur in lecturers' treatment of non-native and native speaker essays - specifically, the conflict that lecturers experience in facing the question of 'same or different' standards. Ballard and Clanchy (1991) maintain that to expect all students' written work to be assessed against the same criteria, is to ignore the qualitative differences between the problems found in work by non-native speakers compared with native speakers. For the lecturers in this study, such an expectation is probably not helping matters. In fact, the findings suggest that there is a potential conflict between the reality of lecturers' treatment of non-native speaker essays, and the view that a common standard that should be used to evaluate all students' work, irrespective of language background. In addition to some of the individual comments that have already been discussed in section 3 (above), this conflict can be seen in the observation that, of the lecturers (fifteen) who reported in Stage 1 that they use the same criteria for both non-native and native speaker essays, most (twelve) also reported that they experience difficulties trying to do this. While the nature of these difficulties was not explored in this study, lecturers' comments provide at least preliminary insights. For example, lecturer 4 explains: "Because the flow of the essay is often difficult to follow, I find it more difficult to apply the standard marking guide to". Lecturer 6 describes the feeling of "fumbling in the dark" when trying to mark non-native speaker essays without guidance on whether or not it is appropriate to use different standards on some criteria. In fact, 'different standards' are already in operation, in so far as the lecturers in this study tend to treat some criteria as less important for non-native speaker essays, and leniency is not uncommon. At the same time though, a sense of discomfort about this has been gleaned from lecturers' comments throughout.

4.2 Data across Stages 1 and 2: Strengths and weaknesses of the methodology

In comparing the two data sets in this study, it was possible (as in some of the examples discussed above in section 3.3.2) to find instances where responses described by lecturers in the questionnaire could be seen to be borne out in practice (as reported in the

protocols). Of course, inconsistencies in the data from Stage 1 to Stage 2 are also of interest for what they might imply about: firstly, the level of awareness lecturers have about their own assessment criteria i.e. are lecturers' own criteria salient enough to them to be reported accurately?; secondly, the difference between the questionnaire data and verbal protocol data i.e. what is the relationship between what lecturers (and raters in general) 'think' about how they assess writing, and what they 'do' when they assess writing?; and finally, the consistency with which each lecturer responds to essays i.e. intra-'rater' consistency.

The data collected in this survey can shed no light on the final of these three issues. Regarding the first however, there are indications that lecturers' own criteria are not entirely salient to them. For instance, like the raters studied by Johns (1991) and Bridgeman and Carlson (1983) whose judgments were influenced by generic expectations about writing style, this feature was rated by most lecturers in the current study as a very important criterion in their own marking schemes. Indeed, lecturer 17, lamenting a consistent lack of "literary style" in his students' essays, remarked (Stage 1): "If I was to mark in terms of style, 90% of students would fail". Yet despite the apparent importance of style, when asked to supply a list of their own assessment criteria, style was not cited by any of the lecturers. This points to an unanswered question of whether lecturers consciously choose not to mark essays against criteria that they really believe to be important (as in the case of lecturer 17), or whether lecturers' knowledge of how they evaluate essays—or what they are able to report about it—is limited. Certainly, Leki (1995) for instance, who found that staff reactions to ESL writing were highly diverse, concluded that ultimately, many were not sure whether writing samples met their criteria or not.

As for the second issue, discrepancies between lecturers' attitudes or beliefs about the way they mark essays (questionnaire responses) and some of the instances of this behaviour (as it is recorded in the verbal protocols) suggest that this relationship is not predictable. For instance, evaluations of grammar/sentence structure occurred in the protocols from physiotherapy and education (see Table 2), yet when asked in the questionnaire to list their assessment criteria, this feature was not cited by the lecturers involved. While it is not possible to conclude anything from this example, it is nonetheless interesting to speculate about what it might indicate: that lecturers are not able to

report their own criteria accurately; that individual lecturers are inconsistent in what they respond to from essay to essay; or both of these. Of course, the overriding consideration here is the fact that a verbal protocol is never an utterly comprehensive record of what happens when a task is performed. As Weigle points out, "the absence of any particular phenomenon in a protocol is not evidence of its absence in actuality" (1994: 207). And just as importantly, as reports of 'one off' instances of behaviour, the generalisability of anything observed in a verbal protocol is low (in the present study, the amount and type of data in the protocols depended as much on each lecturer's own criteria, as it did on the features of the particular essays that were marked).

5. Conclusion

In terms of understanding raters' responses to writing, this study has illustrated the confounding of two issues: i) accuracy of raters' reports about what they attend to when making judgments about writing, with ii) intra-rater consistency; a problem already widely observed in the literature on language testing. In so far as discipline-specialist 'raters', who are not language-trained, need to work out how to respond to non-native speaker essays, this study has also shown, like Vaughan, that "each rater comes to rely on his own method" (1991: 121). In the context of the current study, that lecturers are experiencing uncertainty, or worse, a lack of confidence in their 'own methods', is a problem for the institutions of higher education that are embracing a culture of 'internationalisation'. This is manifest in several ways—for instance, in efforts to make curricula more internationally relevant, in the implimentation of policies on cultural diversity, and in an increasing reliance on fee-paying overseas students—all of which contribute to a pressing need for administrators, educators and researchers to consider the issues faced by academic communities who now require a level of intercultural 'literacy' to be able to function effectively in culturally diverse universities. A particular issue raised by this study, and which warrants such consideration, is that of the implications of a tension between the desirability of common evaluation standards, and the practical necessity of responding in specific ways to non-native speaker essays.

The lecturers in this survey responded with overwhelming support to the suggestion of receiving formalised advice on how they should

respond to non-native speaker essays in ways that are appropriate pedagogically, and equitable. But clearly, further research is needed to determine the nature of any such support. The work of Hawthorne (1997) for example, has already identified some of the requirements of those teaching in culturally diverse settings in terms of cross cultural skills. Similar research is needed to identify what kind of training and/or support would be appropriate for lecturers making assessment decisions. Given that assessment criteria are not always salient to those who use them, continuing to investigate the writing features attended to, as well as other aspects of the whole assessment context that influence judgments remains important (see Hamp-Lyons, 1990). Although possible effects (such as slowing or altering of normal task performance) have already been acknowledged as a limitation of 'think aloud' methods, one lecturer's comment after recording a verbal protocol in this study, is testimony to the value of these methods for improving self-awareness in the marking process:

> *I found it useful because it forced me to articulate what I do and care about. [12]*

Perhaps a bigger issue arising from this study is that of whether or not striving for uniformity of standards is, in fact, desirable. Indeed, two lecturers in Stage 1 commented:

> *Often difficult to mark students with ESL, don't wish to penalise due to language skills – however, good written/verbal language skills essential in physiotherapy profession. [2]*

> *Try to ignore English; Students are here to learn specialist dentistry. [8]*

Both comments, while representing opposing positions in one sense, serve to highlight the importance of the purpose of assessment, or what it is that needs to be measured in any given setting. For the two lecturers above, the expected educational outcomes, or the skills they expect their students to have acquired by the time they graduate, are not the same. Clearly, consistency in standards of assessment is not the only thing at stake. And yet, being 'international' probably means more than instituting acceptable ways of 'making allowances'. Rather, a more profound engagement with difference and diversity would require institutions to ask questions like: which competencies *need* to be mandatory for all graduates, and which do not?

Acknowledgements

This article is a version of a paper presented at the Language Testing Research Colloquium, Vancouver, BC, March 2000. It is based on part of my unpublished Masters dissertation, *Assessment of student essays: Methods of marking work written by students from non-English speaking backgrounds*, The University of Melbourne, 1999, completed under the expert supervision of Catherine Elder. I would like to thank Brian Lynch for reading an earlier draft of this article, and gratefully acknowledge the suggestions he made which led to improvements on the final draft.

6. References

Ballard, B. A. 1993. *English language assistance and other forms of academic support for overseas graduate students.* Occasional Paper GS93/2, Overseas Students Committee, The Graduate School, The Australian National University.

Ballard, B. & Clanchy, J. 1991. Assessment by misconception: Cultural influences and intellectual traditions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 19–35). Norwood, NJ: Ablex Publishing Corporation.

Bell, J. 1993. *Doing your research project; A guide for first time researchers in education and social science (2nd ed.).* Buckingham, MK: Open University Press.

Berkenkotter, C. 1983. Decisions and revisions: The planning strategies of a publishing writer. *College Composition and Communication, 34* : 156–169.

Breland, H. M. & Jones, R. J. 1984. Perceptions of writing skills. *Written Communication, 1,* 1 : 101–119.

Bridgeman, B. & Carlson, S. 1983. *Survey of academic writing tasks required of graduate and undergraduate foreign students* (TOEFL Research Reports, 15). Princeton, NJ: Educational Testing Service.

Brown, A. 2000. An investigation of the rating process in the IELTS Speaking Module. In R. Tulloh (Ed.), *Research Reports 2000*, Vol. 3 (pp. 49-85). Sydney: ELICOS.

Brown, A., Elder, C.A., Iwashita, N. , McNamara, T. & O'Hagan, S. forthcoming. Investigating raters' orientations in specific-purpose task-based oral assessment.

Craswell, G. 1992. *International graduate coursework students and the urgency of adaption to new learning strategies.* Occasional Paper GS92/2, The Graduate School, The Australian National University.

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

Elder, C. 1992. How do subject specialists construe second language proficiency? *Melbourne Papers in Language Testing, 1*, 1: 17–33.

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

Flower, L. & Hayes, J. R. 1980. The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication, 31* : 21–32.

Freedman, S. W. 1979. Why do teachers give the grades they do? *College Composition, 30*, 2 : 161–164.

Geisler, C. 1994. *Academic literacy and the nature of expertise: Reading, writing and knowing in academic philosophy.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, A. 1998. *Verbal protocol analysis in language testing research: A handbook.* Cambridge: Cambridge University Local Examinations Syndicate and Cambridge University Press.

Hamp-Lyons, L. 1990. Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69–87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. 1991. Reconstructing 'academic writing proficiency'. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 127–153). Norwood, NJ: Ablex Publishing Corporation.

Hawthorne, L. 1997. The issue of racial cleavage on campus. Paper presented at the 8th International Student Advisors Network of Australia (ISANA) Conference, Melbourne.

Horowitz, D. 1986. What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20* : 445–462.

Huot, B. 1990. Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41,* 2 : 201–213.

Huot, B. 199a. The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 2 : 237–263.

Johns, A. 1991. Faculty assessment of ESL student literacy skills: Implications for writing assessment. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 167–179). Norwood, NJ: Ablex Publishing Corporation.

Leki, I. 1995. Good writing: I know it when I see it. In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 23–46). Norwood, NJ: Ablex Publishing Corporation.

Meiron, B.E. 1998. Rating oral proficiency tests: a triangulated study of rater thought processes. Unpublished MA thesis: UCLA.

Milanovic, M., Saville, N. & Shuhong, 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 Colloquium, Cambridge and Arnhem* (pp. 92–114). Cambridge: Cambridge University Local Examinations Syndicate and Cambridge University Press.

Miles, M. B. & Huberman, A. M. 1994. *Qualitative data analysis.* Thousand Oaks, CA: Sage Publications.

Nisbett, R. E. & Wilson, T. D. 1977. Telling more than we can know: Verbal reports on mental processes *Psychological Review, 84,* 3 : 231–259.

O'Loughlin, K. 1992. *The assessment of writing by English and ESL teachers.* MA thesis: The University of Melbourne.

Olson, G. M., Duffy, S. A. & Mack, R. L. 1984. Thinking-Out-Loud as a method for studying real-time comprehension processes. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 253–286). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Perkins, D. N. 1981. *The mind's best work.* Cambridge, Mass: Harvard University Press.

Pressley, M. & Afflerbach, P. 1995. *Verbal protocols of reading: The nature of constructively responsive reading.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Raforth, B. A. & Rubin, D. L. 1984. The impact of content and mechanics on judgements of writing quality. *Written Communication, 1,* 4 : 446–458.

Raimes, A. 1985. What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly, 19,* 2 : 229–258.

Samuelowicz, K. 1987. Learning problems of overseas students: Two sides of a story. *Higher Education Research and Development,* 6, 2 : 121–133.

Santos, T. 1988. Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 21–22* : 69–90.

Strauss, A. & Corbin, J. M. 1990. Basics of quantitative research: Grounded theory procedures and techniques. Newbury Park, CAL: Sage Publications.

Swarts, H., Flower, L. & Hayes, J. R. 1984. Designing protocol studies of the writing process: An introduction. In R. Beach & L. S. Bridwell (Eds.), *New directions in composition research* (pp. 53–71). New York, NY: The Guilford Press.

Vann, R. J., Lorenz, F. O. & Meyer, D. M. 1991. Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 181–195). Norwood, NJ: Ablex Publishing Corporation.

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

Weigle, S. W. 1994. Effects of training on raters of ESL compositions. *Language Testing, 11*, 2 : 197–223.

Wyatt, D., Pressley, M., El-Dinary, P. B., Stein, S., Evans, P. & Brown, R. 1993. Comprehension strategies, worth and credibility monitoring, and evaluations: Cold and hot cognition when experts read professional articles that are important to them. *Learning and Individual Differences, 5* : 49–72.

Appendix A

Questionnaire items

1. Do you use a marking guide or marking sheet which shows the assessment criteria for marking essays in your course? NO/YES If 'Yes', please explain where this marking guide comes from, e.g. is it supplied by your department; have you developed your own; do you usually modify the marking guide to include your own criteria?
2. Please describe the assessment criteria you use, i.e. describe what you look for when marking an essay (this might include positive and negative features of writing).
3. Do you usually rank order essays in some way to help you decide on a grade for each one, e.g. place them in order from 'best to worst'; put them into groups such as 'good', 'fair', 'poor'; rate each one against benchmark essays? NO/YES If 'Yes', please describe.
4. Do you apply the same assessment criteria to essays written by non-native speaking students as you do to essays by students who are native speakers of English? YES/NO If 'No', please describe how the criteria differ for marking essays by non-native spekaing students.
5. Do you find it necessary to be more lenient (or more strict) with any of your assessment criteria for essays by non-native spekaing students? NO/YES If 'Yes', please describe.
6. Do you ever experience difficulties in applying the same criteria to NESB student essays as you do to essays by students who are native speakers of English? NO/YES If 'Yes', please describe.
7. Have you or your department implemented any special procedures for marking essays written by non-native spekaing students to assist you in the marking process (this could include guidelines, policy, training, moderation sessions or meetings with your department to discuss marking)? NO/YES If 'Yes', please describe. Do you find this helpful? NO/YES Please explain why/why not.

8. Would you welcome any special procedures or guidelines to assist you in marking non-native speaking student essays? NO/YES Please explain why/why not.

9. When marking essays by students who are **native speakers** of English, how important are each of the following to you? Please circle a number on the scale from 1 ('very important') to 5 ('not at all important'):

	Very important ↓				Not at all important ↓
☐ grammar and sentence structure	1	2	3	4	5
☐ spelling	1	2	3	4	5
☐ punctuation	1	2	3	4	5
☐ vocabulary	1	2	3	4	5
☐ overall structure and organisation; paragraphing; development of ideas	1	2	3	4	5
☐ content (e.g. support for argument; research; relevance to the topic)	1	2	3	4	5
☐ style: is it appropriate to the audience; does it meet the expectations of writing in your academic field profession?	1	2	3	4	5
☐ other (e.g. length, presentation, is it interesting to read?) Please specify:	1	2	3	4	5

10. Now please go back and indicate which THREE features are most important to you by numbering the boxes on the left. Please select three (3) features: number 1 is the most important, number 2 is the second most important, and number 3 is the third most important feature.

11. When marking essays by **non-native speaker students**, how important are each of the following to you? Please circle a number on the scale from 1 ('very important') to 5 ('not at all important'):

	Very important ↓				Not at all important ↓
☐ grammar and sentence structure	1	2	3	4	5
☐ spelling	1	2	3	4	5
☐ punctuation	1	2	3	4	5
☐ vocabulary	1	2	3	4	5
☐ overall structure and organisation; paragraphing; development of ideas	1	2	3	4	5
☐ content (e.g. support for argument; research; relevance to the topic)	1	2	3	4	5
☐ style: is it appropriate to the audience; does it meet the expectations of writing in your academic field profession?	1	2	3	4	5
☐ other (e.g. length, presentation, is it interesting to read?) Please specify	1	2	3	4	5

12. Now please go back and indicate which THREE features are most important to you by numbering the boxes on the left. Please select three (3) features: number 1 is the most important, number 2 is the second most important, and number 3 is the third most important feature.

13. What kind of feedback do you give students about their performance? (You may tick more than one):

☐ mark/grade only

☐ a copy of your marking sheet

☐ another kind of feedback (e.g. margin notes; a special feedback sheet designed for giving to students) Please specify

Appendix B

Instructions for recording 'think aloud' protocols *

1. Begin by turning on the tape recorder and saying the time and the date. Replay it to make sure the recorder is working. Identify each new essay as you begin, e.g. essay #1; state the topic and word length.
2. Say whatever's on your mind. Don't hold back vague thoughts or undeveloped ideas.
3. Speak as continuously as possible. Try to say something at least once every five seconds.
4. Speak audibly. Watch out for your voice dropping as you become involved.
5. Speak as you would just for yourself, not for an audience. Don't worry about complete sentences or being eloquent.
6. Get into the pattern of saying what you're thinking now - while you are reading and marking an essay - *not when you have finished.*
7. When you have finished the session, please say something like, "This is the end of my session for today" followed by the time and the date.

* Based on: Perkins, D. N. 1981. *The mind's best work.* Cambridge, Mass: Harvard University Press; Geisler, C. 1994. *Academic literacy and the nature of expertise: Reading, writing and knowing in academic philosophy.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Written Prompts

- What are you doing now? (skimming whole essay? scanning for particular information? reading introduction/conclusion/which paragraph/section?)
- What are you looking for in this essay?
- What are you noticing about this essay as you read?
- Are you thinking about other essays you've read?
- How are you reacting to what you are reading now?
- Are you writing comments/corrections? what are you writing?
- How are you deciding what mark you will give this essay?

Appendix C
Coding schedule for verbal protocols

Frequency

ANALYTIC EVALUATIONS

Content
good ideas: good; correct; good level of detail; good understanding of course content
good/satisfactory discussion: adequate level of argument; good support for ideas; well
satisfactory: includes some required points/key words; ideas adequate/reasonable
weak
not accurate: inadequate/incomplete understanding of topic
not relevant: not central to topic; relevance unclear; doesn't address topic
not enough: doesn't include everything; doesn't say much
unsatisfactory discussion: not sufficiently developed; lack of explanation; superficial
problems with development of argument: interference from student's first language
inadequate synthesis of source information: straight from lectures

Structure/organisation
introduction good/as required
introduction missing/inadequate
good: ideas well sign-posted; good paragraph structure; logical organisation of ideas
problems making comprehension difficult: lacks planning, headings, coherence, logical sequencing, cohesion btwn sections and whole; sections not linked to topic
succinct presentation of ideas: comes to the point, even if a bit short
lacks succinctness: long-winded; rambles on; not specific enough
repetition of ideas/redundancy
evidence of effort to make a point, although long-winded style

Grammar/sentence structure
no grammatical errors
grammatical errors
poor: makes comprehension of ideas difficult
characteristic non-native speaker errors

Overall writing ability
good: reads well; good English; well written, despite small problems with coherence
English shows improvement since previous work; may have sought ESL assistance
good English but numerous grammatical errors
doesn't read well

Referencing
accurate/as required
not accurate/errors
missing: unacknowledged material used verbatim

Vocabulary
vocabulary problems: incorrect; inappropriate; confusion with another word

Length
too long
too short

Other
register problems: tone too colloquial; lacks consistency
legibility poor: bad writing

continued...

...continued

READING STRATEGIES

scans for key words - poor grammar/organisation (makes ideas hard to understand)
puts extra effort into reading, overlooking grammar - poor grammar but good content
re-reads sections - confusion between current essay and those read earlier
guesses - poor legibility

AFFECTIVE RESPONSES

enjoy reading: work is well written, well organised
pleasing: good content showing student has learnt from the course
a relief: good support for ideas/well argued
amusing/ironic: connotations of vocabulary error
evocative of reader's research interests: gaps in student's intercultural knowledge
distracting: influence of student's first language into English (errors)
annoying: grammatical errors
annoying: poor structure & organisation; long-winded writing
tiring: poor structure & organisation
frustrating: poor legibility

Yes / No

GRADING/CORRECTIONS/FEEDBACK

awards approximate grade (letter or percentage range; pass/fail)
awards grade based on benchmark; provisional on standard of others yet to be read
makes analytic comparison: compares essay with other essays against various criteria
indicates/corrects small grammatical error
indicates error, but doesn't correct, where lengthy correction required
corrects reference
ticks good idea
marks inadequate explanation with query
corrects/marks vocabulary error
crosses out redundance
writes margin notes about inaccuracies in content
after scanning, re-reads closely, making annotations
writes feedback addressing formal criteria
shows leniency to failure to credit source
monitors tone of comments when correcting annoying errors

GLOBAL EVALUATIONS

good; well done
shows effort/thought
adequate/reasonable standard
not very good

CONTEXTUAL INFORMATION

reads aloud from student text
describes/paraphrases text: describes/summarises content; states essay topic
identifies/names author (student)
states expectation/belief about student's language background
describes location in essay/reports on reading progress
incidental comments (e.g. microphone, background noise/interruptions)

Appendix D

Descriptive statistics of lecturers' ratings of writing features for their importance as assessment criteria

Shown below are range, mean, and median scores for ratings on a 5–point scale from 'very important' (score = 1) to 'not at all important' (score = 5):

1. for native speaker essays

		content	struct/org	style	grammar/ss	spelling	vocabulary	punctuation
All Depts	min	1	1	1	1	1	1	1
(*N* = 21)	max	2	2	4	5	5	4	5
	mean	1.048	1.333	2.095	2.238	2.524	2.619	2.833
	median	1	1	2	2	2	3	3
Dental Sc	min	1	1	1	1	1	1	1
(*n* = 9)	max	2	2	4	5	5	4	5
	mean	1.111	1.556	2.333	2.778	2.889	2.556	2.889
	median	1	2	2	2	3	2	3
Physioth	min	1	1	1	1	1	1	2
(*n* = 9)	max	1	2	3	3	4	4	4
	mean	1.000	1.222	1.889	2.000	2.222	2.556	2.833
	median	1	1	2	2	2	3	3
Education	min	1	1	2	1	2	3	2
(*n* = 3)	max	1	1	2	2	3	3	3
	mean	1.000	1.000	2.000	1.333	2.333	3.000	2.667
	median	1	1	2	1	2	3	3

2. for non-native speaker essays

		content	struct/org	style	grammar/ss	spelling	vocabulary	punctuation
All Depts	min	1	1	1	2	2	1	2
(*N* = 21)	max	2	3	4	5	5	5	5
	mean	1.048	1.381	2.286	3.000	3.143	2.952	3.238
	median	1	1	2	3	3	3	3
Dental Sc	min	1	1	1	2	2	1	2
(*n* = 9)	max	2	3	4	5	5	4	5
	mean	1.111	1.556	2.667	3.222	3.222	2.778	3.222
	median	1	1	3	3	3	3	3
Physioth	min	1	1	1	2	2	2	2
(*n* = 9)	max	1	2	3	5	5	5	5
	mean	1.000	1.333	1.889	3.000	3.333	3.222	3.556
	median	1	1	2	3	3	3	3
Education	min	1	1	2	2	2	2	2
(*n* = 3)	max	1	1	3	3	3	3	3
	mean	1.000	1.000	2.333	2.333	2.333	2.667	2.333
	median	1	1	2	2	2	3	2

'min' - minimum; 'max' - maximum; struct/org – structure/organisation; grammar/ss – grammar/sentence structure