EXPLORING THE ESTABLISHMENT OF SCORING CRITERIA FOR WRITING ABILITY IN A SECOND LANGUAGE: THE INFLUENCE OF BACKGROUND FACTORS ON VARIABILITY IN THE DECISION-MAKING PROCESSES OF FOUR EXPERIENCED RATERS OF ESL COMPOSITIONS

by

Muhammad Usman Erdosy

A thesis submitted in conformity with the requirements for the degree of Master of Arts
Department of Curriculum, Teaching, and Learning
Ontario Institute for Studies in Education of the
University of Toronto

© Copyright by Muhammad Usman Erdosy (2000)
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-49745-3
Abstract

Exploring the establishment of scoring criteria for writing ability in a second language: the influence of background factors on variability in the decision-making processes of four experienced raters of ESL compositions.

Muhammad Usman Erdosy

Master of Arts, 2000

Department of Curriculum, Teaching and Learning;
Ontario Institute for Studies in Education
University of Toronto

I examined the verbal protocols 4 raters of ESL compositions provided while rating 60 TOEFL essays in an experimental context, where they were instructed not to rely on the usual scoring rubric for this test. I subsequently interviewed the raters to discover how they constructed scoring criteria and how background factors influenced this process. The data revealed that all 4 attempted to infer proficiency from performance by placing writers on a developmental trajectory they established for language learners, based largely on their teaching experiences. To establish specific rating criteria, raters combined their ideas of learners' development with information regarding scoring guidelines, test use and test takers, and relied on assessment experience to develop reading and rating strategies. Differences in scoring criteria emerged through differences in raters' backgrounds, particularly in their teaching experiences, which led them to differ in identifying the key competencies learners mastered at various levels of development.
Contents

List of tables v
List of figures vi
List of appendices viii

Chapter 1. Introduction and research questions 1
  1.1 Research questions 2

Chapter 2. Literature review 6
  2.1 Key background factors influencing raters 8
    2.1.1 Cultural background 8
    2.1.2 Linguistic background 10
    2.1.3 Academic background 11
    2.1.4 Professional experience 13
  2.2 Implications 15

Chapter 3. Research design 17
  3.1 Participants 17
  3.2 Sources of data and instrumentation 20
    3.2.1 Concurrent verbal protocols 21
    3.2.2 Questionnaires 24
    3.2.3 Interviews 24
  3.3 Methods of analysis 29
  3.4 Construction of the case studies 32

Chapter 4. Sam 35
  4.1 Research question A 35
  4.2 Research question B 38
  4.3 Research question C 41
  4.4 Research question D 42
  4.5 Summary 52

Chapter 5. Chris 54
  5.1 Research question A 54
  5.2 Research question B 57
  5.3 Research question C 59
  5.4 Research question D 62
  5.5 Summary 69

Chapter 6. Alex 71
  6.1 Research question A 71
  6.2 Research question B 73
  6.3 Research question C 75
  6.4 Research question D 79
List of tables

Table 1: Principal characteristics of participants 19
Table 2: Coding scheme for decisions made while rating ESL compositions 30
Table 3: The structure of the case studies 33
List of figures

Figure 1: Relative frequencies of interpretive strategies during Sam's rating session 44
Figure 2: Relative frequencies of judgemental strategies during Sam's rating session 44
Figure 3: Relative frequencies of interpretive strategies during phase II of Sam’s rating session 45
Figure 4: Relative frequencies of judgemental strategies during phase II of Sam’s rating session 45
Figure 5: Relative frequencies of interpretive strategies during phase III of Sam’s rating session 47
Figure 6: Relative frequencies of judgemental strategies during phase III of Sam's rating session 47
Figure 7: Relative frequencies of interpretive strategies during Chris’ rating session 63
Figure 8: Relative frequencies of judgemental strategies during Chris’ rating session 63
Figure 9: Relative frequencies of interpretive strategies during phase II of Chris’ rating session 64
Figure 10: Relative frequencies of judgemental strategies during phase II of Chris’ rating session 64
Figure 11: Relative frequencies of interpretive strategies during phase III of Chris’ rating session 65
Figure 12: Relative frequencies of judgemental strategies during phase III of Chris’ rating session 65
Figure 13: Relative frequencies of interpretive strategies during Alex’s rating session 80
Figure 14: Relative frequencies of judgemental strategies during Alex’s rating session 80
Figure 15: Relative frequencies of interpretive strategies during phase II of Alex’s rating session 81
Figure 16: Relative frequencies of judgemental strategies during phase II of Alex’s rating session 81
Figure 17: Relative frequencies of interpretive strategies during phase III of Alex’s rating session 82
Figure 18: Relative frequencies of judgemental strategies during phase III of Alex’s rating session

Figure 19: Relative frequencies of interpretive strategies during Jo’s rating session

Figure 20: Relative frequencies of judgemental strategies during Jo’s rating session

Figure 21: Relative frequencies of interpretive strategies during phase II of Jo’s rating session

Figure 2: Relative frequencies of judgemental strategies during phase II of Jo’s rating session

Figure 23: Relative frequencies of interpretive strategies during phase III of Jo’s rating session

Figure 24: Relative frequencies of judgemental strategies during phase III of Jo’s rating session

Figure 25: Relative frequencies of interpretive strategies during phase IV of Jo’s rating session

Figure 26: Relative frequencies of judgemental strategies during phase IV of Jo’s rating session
List of Appendices

Appendix A: Letter of consent signed by participants in the study 124
Appendix B: Conventions used in transcribing concurrent verbal protocols 128
Appendix C: Sample transcribed protocol 129
Chapter 1: Introduction and research questions

Although direct assessment has become the method of choice in judging writing ability in English as a Second Language (ESL), and impressive levels of inter-rater reliability have been achieved by training raters in the use of specific rating scales (Weigle, 1994), lack of agreement in defining the construct which direct writing assessment aims to measure remains a concern for validation studies (Milanovic, Saville & Shuhong 1995, p. 93). One potentially fruitful response to this concern is represented by efforts to link variability in raters' judgements of writing ability to corresponding sources of variability in their backgrounds. Unfortunately, however, such studies have focused on isolated factors, relating, for example, level of training (Weigle, 1994) or linguistic background (Brown, 1995) to patterns of decision making, themselves frequently reduced to a single dimension such as level of severity. Even Pula and Huot's (1993) study, which modeled a range of influences on the behaviour of raters of English compositions through the useful concept of discourse communities, neglected such factors as mother tongue and cultural background, which probably play a significant role in the assessment of ESL compositions, given the diversity found among teachers and assessors of ESL. Consequently, there is, at present, no satisfactory explanation of how raters of ESL compositions construct rating criteria, even if numerous formative influences on their strategies and judgements have been tentatively identified.

It is in this context that I designed the present study, which I conducted in two stages. First I analyzed data generated through questionnaires and think-aloud protocols by a project investigating raters' decision-making behaviours (Cumming, Kantor, & Powers, 1998) where raters were asked to assess 60 compositions without relying on a rating scale. My aim was to provide detailed descriptions of the way four experienced raters rated ESL compositions in the absence of guidelines, a condition which may not be reflective of standard assessment practice but which the investigators of the Cumming et al. (1998) study felt would reveal the complexity of the process of constructing scoring criteria. Subsequently, I explored the influ-
ence of background factors on this process through interviews with the participants, which were designed based on my analysis of the think-aloud protocols. This part of my study was conducted independently and aimed at contributing to the construct validation of direct writing assessment through what may be termed an interpretive case study: "case study," because its aim was to find out what participants think, rather than to judge their opinions; "interpretive," because "the descriptive data gathered [was] used to develop conceptual categories [and] illustrate, support, or challenge theoretical assumptions held prior to data gathering" (Merriam, 1988, 27-28).

In designing my study I limited its scope in a number of ways. To begin with, I took the usefulness of direct writing assessment for granted; my aim was to help establish its validity, rather than to weigh its merits against other methods. Similarly, I accepted the value of large-scale, standardized testing in spite of searching criticism (White, 1994; Williamson, 1993), and limited my examination of rater attitudes and behaviours to this context. Finally, although I was not explicitly concerned with rater training, I acknowledged that exploring improved training methods would constitute a logical extension to my study relating raters' construction of rating criteria to background factors. In spite of these limitations, the present study should assist in the validation of direct writing assessment for two reasons: it takes a multi-faceted view of both the processes of constructing scoring criteria and the background factors that influence raters, and, rather than offering generalizations about groups of raters, it focuses on how individual raters with diverse backgrounds render specific judgements in a specific rating situation.

1.1 Research questions

Aiming to complement Cumming, et al's (1998) study concerning the decision-making behaviours of raters of ESL compositions, by addressing an issue that was not explored there in detail, I initiated my enquiry with the following question:
What processes do raters of ESL compositions follow in constructing scoring criteria to assess a corpus of compositions, and how can these processes be related to their personal and professional backgrounds?

In order to sharpen my focus, I then followed Kroll (1998, p. 223) in identifying "reader," "written text," "writer," "writing prompt," and "scoring procedure" as the critical variables in direct writing assessment. Using this model, which views raters as readers bringing specific purposes and expectations to the rating task (Huot, 1990; Kroll, 1998; Pula & Huot, 1993; Purves, 1984; White, 1994; Williamson, 1993), I was able to generate four research questions — exploring raters' attitudes to, in turn, the written text, the writer, the writing prompt and the scoring procedure — which I felt covered the complexity of the process of constructing scoring criteria.

Such questions provided a useful first approximation of the processes raters in the study followed in constructing scoring criteria, but needed refinement upon analyzing the concurrent think-aloud protocols generated in the first part of the study. Above all, I found that my participants' attitudes towards a written text represented answers to a more fundamental question, namely, "How can performance (in this case embodied in a text written under carefully specified constraints) be related to proficiency?" Answering this question constituted my participants' initial step in the process of constructing scoring criteria; consequently, I phrased my first specific enquiry as follows:

Research question A: In what principled way did the raters of ESL compositions participating in the study relate the performance embodied in a written text to a writer's level of proficiency, and how did background factors influence this process?

Next, examining the perceptions of my participants concerning the writers of the compositions they were rating, I found that they manifested themselves in two ways. On the one hand, the participants were concerned with the impact of situational factors (such as fatigue or lack of time) on performance, and I could subsume this aspect of their behaviour under the question addressing their views on the meaning of a performance. On the other hand, the participants were aware of potential handicaps arising from a mismatch between writers'
cultural backgrounds and certain essay topics. In turn, such a concern could be subsumed under a research question addressing raters' attitudes to writing prompts (see Kroll & Reid, 1994) in general. In addition, interpreting the role of writing prompts involved assumptions about test use and test takers, which were also included in the discussion of the second research question, which was eventually articulated in the following manner:

*Research question B:* How did raters of ESL compositions interpret the role of writing prompts in direct assessment, and how did background factors influence this process?

In designing the third question, I followed Kroll (1998, p. 223) in breaking down "scoring procedure" into two distinct aspects: reading strategies and rating criteria. Although reading strategies had no direct bearing on the construction of rating criteria, each participant's rating session in the present study was structured in distinct stages resulting from their approaches to reading both individual compositions and the entire corpus of compositions, and it was impossible to describe their behaviour without the contexts provided by these stages in their procedures. Consequently, I decided to address this issue in the third research question:

*Research question C:* What reading strategies did raters of ESL compositions establish (to deal with both individual compositions and a corpus of compositions) and how did background factors influence these?

Finally, in formulating the fourth research question, I aimed at presenting raters' scoring criteria as the product of the interaction between raters' assumptions concerning the meaning of performance, as well as their information on test use and candidate population. Even in situations where raters are presented with a scoring rubric, such additional factors play an important role in the application of scoring criteria to compositions; in the present study, where raters were asked to construct their own rating criteria they played an even more crucial role. Such considerations produced the concluding research question, which presented rating criteria as the outcomes of a complex process:
Research question D: By what process did raters of ESL compositions construct specific rating criteria, and how did background factors influence this process?

To go beyond the insights provided by earlier studies, I felt that it was important to review the published literature concerning the influence exerted by backgrounds factors on the decision-making behaviours of raters of ESL compositions. The results of this literature review are presented in Chapter 2. Subsequently, in Chapter 3, I describe the research design I employed in response to my research questions and the literature review, including my choice of participants, research instruments, and methods of analysis. In Chapters 4 to 7 I then present the case studies of the four participants, including the steps they followed in constructing scoring procedures and scoring criteria and the influence of background factors on this process, based on my analyses of their think-aloud protocols, questionnaires, and interviews. Finally, in Chapter 8 I answer the research questions originally posed and discuss the implications of the study's findings.
Chapter 2: Literature review

Each of my research questions discussed in Chapter 1 presented two aspects: on the one hand, I was concerned with identifying the process raters of ESL compositions followed in constructing scoring criteria; on the other hand, I wished to specify the influence of background factors on the steps involved in that process. In exploring the first aspect, I found that I could use existing models of the rating process (particularly, Kroll, 1998) as well as existing categorizations of rater behaviour (particularly Cumming, 1990 and Cumming, et al., 1998) in constructing my hypotheses. However, I found explanations relating background factors to the behaviour of raters of ESL compositions unsatisfactory because, as discussed in Chapter 1, they reduced not only background factors but also the complex process of constructing, or applying, scoring criteria to single dimensions. At the same time, several background factors have been identified in the literature as potentially significant. Consequently, I decided to set the context for addressing the second aspect of my research questions by critically examining current hypotheses concerning the interaction of background factors in shaping the attitudes and behaviours of raters of ESL compositions. Such a literature review could also assist me in selecting participants for my study, and in probing them through interviews, by identifying potentially significant aspects of their backgrounds. In compiling it, I focused on publications concerned with the raters of ESL compositions, although I did not hesitate to refer to studies concerning the evaluation of compositions written by native speakers (NS) of English, of oral proficiency, as well as of proficiency in other languages, if these provided relevant insights.

Concerning methodology, perhaps the most fruitful approach to studying the attitudes and behaviour of raters of second language compositions, in view of the sheer quantity of data generated, has been to elicit verbal protocols (Connor & Carrell, 1993; Cumming, 1990; Vaughn, 1991), supplemented in some instances by written reports (Milanovic, Saville & Shuhong, 1995) and interviews, during and after rating sessions. Based on such data,
Researchers have classified and counted decision-making behaviours, besides correlating them with scores assigned by raters as well as with independent measures of writing quality (e.g., Cumming & Mellow, 1996; Tedick & Mathison, 1995). Factor analysis of rater judgements has also been used in identifying areas of communicative competence that raters considered significant, with the results correlated with background factors (Chalhoub-Deville, 1995a). In a third type of design (e.g., Santos, 1988), raters were asked to rate compositions on a multi-trait scale, so that overall scores could be correlated with scores on individual scales, in order to see how a rater's background might determine the areas of language he/she will focus on in evaluating compositions.

In addition to directly observing rater behaviour, indirect information on rating criteria can be obtained from, for example, rating scale descriptors, which, for better or worse, represent expert consensus involving large and diverse rater populations (see Brindley, 1998 for a recent overview and critique). Similar generalizations concerning rater expectations may be inferred from the prescriptions of writing textbooks (e.g., Oshima & Hogue, 1991), concerning rhetorical features. Finally, cross-cultural comparisons of writing styles are also available (Grabe & Kaplan, 1989, 1996; Kachru, 1988; Kaplan, 1983).

The image I see emerging from these studies is that four factors — culture, mother tongue, academic background, and professional experience — have been shown to exert a particularly strong influence on rating behaviour, although none will be sufficient to explain it in isolation. Before examining their respective roles, however, it is important to note that variability in judgements concerning English compositions had been statistically demonstrated by the early 1960s (see Diederich, French & Carlton, 1961), yet the exploration of background influences on these judgements had to await the 1990s. Such sudden interest in raters' backgrounds is best explained by the redefinition of reading as meaning-making, rather than as accessing information encoded in texts by writers, with attendant shifts in conceptualizing the role of raters, as advocated by Huot (1990), Purves (1984), White (1994), and Williamson (1993) in the context of assessing the English compositions of native speakers, and by Janopoulos
(1993) in the context of assessing ESL compositions. Such a redefinition of reading has been linked, in turn, to "schema-theoretical accounts of reading by Rumelhart (1980); Collins, Brown and Larkin (1980); Spiro (1980) and Anderson (1977)" (Pearson & Tierney 1984, p. 144), although with significant contributions from the discipline of literary criticism inspired by Speech Act Theory (Fish 1980, pp. 197-245). What is crucial here is that if raters of compositions are indeed readers, relying on their experiences, expectations, and purposes to construct meaning, rather than umpires, using absolute standards to evaluate the effectiveness with which writers encode meaning in a text, then background becomes the key to explaining rater variability.

2.1. Key background factors

2.1.1. Cultural Background

Among the background factors influencing raters of ESL compositions, the importance of cultural heritage was perhaps the first to be appreciated, thanks to Kaplan's (1966) assertion that rhetorical patterns varied across cultures. It was following his seminal work that contrastive rhetoric was established as a distinct field of study, essentially extending the contrastive analysis of sentence-level grammar to the area of discourse competence.

The divisions established by Kaplan (e.g., "Oriental," "Semitic," "Romance") appear, in retrospect, to be too broad, and the contrastive study of rhetoric has moved well beyond his original formulation. In addition, the approach is dangerously open to stereotyping (Raimes 1998). However, the existence of culturally determined rhetorical conventions, and their influence on raters' judgements, can certainly be observed in the case of British and North American English, with the stress on linear reasoning being paramount. Indeed, Kaplan (1966) himself advocated contrastive studies principally as a means to designing exercises that would improve learners' control of English rhetorical devices.

---

1 Schema theory, of course, represents only one strand in the broader field of individual difference research in reading (Daneman, 1984; Samuels & Kamil, 1984), even if it permeates the literature on writing assessment.
In view of the expectation of linear reasoning, it is not surprising to find, for example, that a recent survey of holistic writing assessment of ESL compositions (Tedick & Mathison, 1995) yielded a generally positive correlation between rater judgements about the effectiveness of framing, measured by the ability of raters to predict the topic and the pattern of development of an essay after reading only the first paragraph, and the score awarded. Extension of these results to the sphere of academic writing in English as well as ESL is suggested by a study of peer reviews in scholarly journals (Belcher, 1995), which found that a clear thesis statement at the outset was considered essential. At the same time, the magnitude of the task of getting non-native-speaking (NNS) writers to adopt the English rhetorical style may be gauged from a survey by Kobayashi and Rinnert (1996), which revealed that NNS (here, Japanese) learners and to some extent even teachers, preferred the rhetorical patterns of their own culture even in English prose.

Judgements related to content can be similarly affected by cultural considerations. Perhaps the key expectation in Anglo-Saxon contexts is writing that, in Cumming's (1995) words, not only transmits but also transforms knowledge. This attitude is commonly contrasted with, for example, the popularly perceived tendency of Asian writers to recycle, rather than build upon, a body of knowledge, and their corresponding distaste for criticism. Another significant factor, noted by Basham and Kwachka (1991), concerns the Anglo-American preference for detachment, particularly in an academic context, which may be contrasted with other cultures' tendency to personalize even scholarly issues.

When we add to the list preferences involving tone and style, such as using an indirect strategy to present bad news in both academic (Belcher, 1995) and business transactions (Guffey & Nagle, 1997), the importance of relating cultural background to rater judgements and strategies becomes even greater. Yet, it should be stressed that no single background factor can account for rater behaviour. Even attitudes to rhetorical features, which are deeply

---

2 To highlight the dangers of such "Orientalist" stereotyping, a study of classical Hindu legal texts (Lingat, 1972) reveals that side by side with the (admittedly frequent) recycling of knowledge a great deal of innovation occurs, frequently in the face of sanctions. Studies of caste (especially, Dumont, 1970) show a similarly broad capacity for innovation and evolution.
ingrained in culture, can be superseded. For example, data cited by Kobayashi and Rinnert (1996) showed that coherence at the paragraph level was accorded greater significance by all raters in their study, regardless of their cultural backgrounds, than adherence to a particular rhetorical pattern. Tedick and Mathison (1995) likewise found that even the failure to frame a composition could be overlooked in certain circumstances. Thus, while I regard cultural background as significant, I will treat it as only one of a range of factors bearing on rater behaviour.

2.1.2. Linguistic background

Although raters' linguistic background may be subsumed under their cultural heritage, several studies have attempted to capture the contrasting attitudes of NS and NNS raters. Certainly, given the internationalization of English and the increasing number of NNS assessors around the world, this perspective is important. However, few uncontested generalizations have emerged in spite of the amount of research invested. For example, some studies (e.g., Fayer & Krasinski, 1987; Santos, 1988) have found NNS raters to be more severe, others (e.g., Brown, 1995) have found them to be generally more lenient. Those who found NNS to be more severe generally explained their findings by referring to the considerable time and energy NNS assessors had invested in learning a language, which lead them to attribute errors to a lack of commitment on the learners' part (Santos 1988, p. 85). This view finds limited confirmation in Brown's (1995) study of the assessment of the proficiency level of Japanese-speaking tour guides: here NNSs were more lenient in every respect, except in judging the mastery of politeness strategies, a skill that NNSs generally have the greatest difficulty comprehending.

In trying to account for inconsistencies in published research, one may again criticize a preoccupation with a limited set of factors, at the expense of others that may have also influenced behaviour. For example, the NS judges in Santos' study were college professors and their ratings had no practical consequences for the writers since they were used only for research purposes. By contrast, in Brown's study NS judges were not only "naive" assessors, without language teaching and assessment experience, but also had the task of judging whether
the candidates they assessed qualified for membership in their profession based on their proficiency levels; their decisions thus carried serious consequences for applicants and for the reputation of their profession. Under such circumstances it may not be surprising that the first group of NSs in Brown's study turned out to be lenient in comparison with NNSs, while the second turned out to be harsh.

In studying NS-NNS contrasts we must also consider the status of a language and its speakers, which may considerably complicate the picture. Also, it would be a mistake to reduce the differences between NSs' and NNSs' reactions to the single dimension of rater severity. These are two key lessons to be learnt from studies conducted by Chalhoub-Devile (1995a, 1995b) concerning varying perceptions among Arabic speakers of proficiency in Modern Standard Arabic (MSA), the *lingua franca* which enables the speakers of various Arabic dialects to communicate with each other. While in this situation the concepts of NS and NNS do not strictly apply, Chalhoub-Devile found that Arabic speakers living in Lebanon equated learner proficiency in MSA almost exclusively with accuracy, whereas Arabic speakers living in the United States valued overall communicative quality as much as accuracy. The reason for the difference, according to Chalhoub-Devile, lay in the divergence between the Lebanese Arab perception of MSA as a language to be used only in highly formal situations, and the American Arab view that it could be used in a wide range of situations, informal as well as formal.

The overall impression gleaned from all the studies is that raters' linguistic status cannot be reduced to a simple NS-NNS contrast. Moreover, it clearly interacts with other factors, as well as with the context and purpose of the ratings rendered, and research designs have to bring out this interplay in order to explain certain patterns of behaviour.

2.1.3. Academic Background

Given the rapid expansion of ESL programs, attracting many professionals from other fields of academia to language instruction, raters of ESL writing often carry over considerable
expertise, including assumptions and behavioural patterns, to the rating process. In addition, as has just been mentioned, students' writings will be judged in contexts beyond the second language classroom, sometimes by people used to the writing of native speakers. Therefore, academic background is worth including in the list of factors potentially influencing rater judgements.

The central problem, noted by Leki (1995, p. 41), is that while academics may accept in theory that there is no "absolute quality without context," they assume in practice that there is, and that their own standards are the absolute ones. An example of diversity across disciplines may be the contrast between the "cumulative and consensual" approach to knowledge of the natural sciences and the confrontational literary style found in the humanities and the social sciences (Belcher, 1995). There is, likewise, a gap between the expectations of content-area teachers and ESL teachers. To the former, content — developing ideas, providing empirical support, and forming judgements — is usually the key concern. For example, Santos (1988) found that college professors (NS as well as NNS) rated content consistently more severely than language. Although they viewed errors as linguistically unacceptable, as long as sentences remained comprehensible they tolerated errors (though not stigmatized forms such as double negatives), because, according to Santos, they had become resigned to the generally lower proficiency level of their students. By contrast, ESL teachers, who could not take language for granted, frequently regarded content as a vehicle for self-expression and valued linguistic accuracy over factual accuracy or stylistic considerations.

Finally, teachers of English to native speakers have been found to be consistently more severe than ESL teachers in their treatment of sentence-level errors, which may be the only generalization with widespread empirical support in the existing published literature. The differences in judgement may be attributed to cumulative context effects. Research (e.g., Miller & Crocker, 1991; Sweedler-Brown, 1993) has shown that any composition will receive a higher score if preceded by weak compositions than if preceded by strong ones. Consequently, teachers of English to NSs, used to seeing relatively proficient essays, may be
liable to react more negatively to NNS errors than ESL teachers who are used to ESL learners' errors. Nor should we be surprised that teachers of English to NSs may overlook positive discourse features in otherwise weaker essays since there is research to suggest that raters will pay attention to only selected aspects of proficiency at different levels. Pollitt and Murray (1995), for example, have shown that raters pay attention mostly to grammatical competence at the lowest level of proficiency, moving to sociolinguistic competence in the middle levels and discourse competence at the top level. Upshur and Turner (1995) even attempted to set up "empirically-derived, binary-choice, boundary-definition scales" to reflect such a mode of thinking, although the details of their intriguing scheme have yet to be worked out satisfactorily.

2.1.4 Professional experience

The clearest conclusion to emerge in this area is that professional and lay raters differ considerably, though not consistently, in their judgements of error severity. For example Bamwell (1989) found professional raters to be more tolerant of language errors than lay raters were; he suggested, not unreasonably, that their exposure to the widest possible range of linguistic ability enables them to put learners into a more realistic, thus more forgiving, perspective. Others (e.g., Galloway, 1977; Hadden, 1991) have found otherwise. The differences even extend to the various aspects of language ability being judged: thus, Hadden (1991) identified pronunciation as an area where professional raters' judgements were more severe, whereas Brown (1995) concluded the opposite: professional raters in her study were harsher in all areas of proficiency except pronunciation, although their global ratings were comparable.3

Apart from showing the pitfalls of studying aspects of raters' background in isolation, these studies also demonstrate that raters' judgements are likewise impossible to reduce to a single dimension, such as degree of severity. In spite of shortcomings, however, the influence

---

3 Although my thesis explores how raters behave, not how they should behave, the issue of professional background is relevant also because teachers have yet to fully address the question of whose judgements are to take precedence outside the ESL classroom. What is the value, for example, of a balanced assessment of proficiency, if, as in Brown's (1995) study, a candidate's inability to fulfil a vital task renders him inadequate in the eyes of her peers in spite of the general excellence of his/her performance? Although such a question is surely embedded in Messick's (1989) expanded concept of construct validity, it remains to be fully addressed.
of training and experience on the behaviour of raters of ESL compositions has been dealt with abundantly in recent studies, and remains perhaps the most extensively researched aspect of raters' backgrounds. Key findings, in short, include the fact that experienced raters tend to be less influenced by surface features and more capable of examining language use, content and rhetorical organization concurrently (Cumming, 1990). In addition, training, a related concern, can help raters apply specific scoring schemes consistently (Connor & Carrell, 1993; Weigle, 1994). On the other hand, the benefits of training may be short-term (Lumley & McNamara, 1995), and it will still not be possible to eliminate context effects (Hughes, Keeling & Tuck, 1983) such as described above, where a composition preceded by high quality compositions is prone to getting a lower score than one preceded by poor compositions. Finally, as Hamp-Lyons (1990, p. 81) sensibly points out, "The context in which training occurs, the type of training given, the extent to which training is monitored, the extent to which reading is monitored, and the feedback given to readers all play an important part in maintaining both the reliability and the validity of the scoring of essays."

Another key finding, of course, is that even experienced raters will vary in their judgements and behaviour, as any study of inter-rater reliability will show. One reason for such divergent ratings may be raters' familiarity with specific rating scales in assessing compositions, whether in class, or through standardized tests. Consequently, one of the criteria for choosing participants for the present study was whether they made reference to specific rating scales in the questionnaire concerning their professional background that they completed immediately after generating think-aloud protocols. On the whole, just as raters will bring personal experience to rating situations even when asked to use a specific scale, so they may rely on their previous knowledge of specific scales in a situation where there are no scales to guide them.

Finally an additional useful distinction may be concurrent or previous teaching experiences, which raters may be prone to carrying over into the rating process. I was particularly interested in observing if participants with teaching experience in the present study were
focusing their evaluation on textual features that play an important part in their teaching curricula, on the assumption that raters with teaching experiences are accustomed to giving feedback as a means of encouraging learners to learn (Prior 1995). One way to measure this was by my noting the frequency with which participants edited texts and classified errors in their rating sessions.

2.2 Implications

My key conclusion based on the published literature is that there is no point in perpetuating an atomistic approach that seeks to explain rater behaviour in terms of the influence of individual factors, since neither raters' backgrounds and nor their judgements are reducible to a single dimension. At the same time, there are few useful models relating a range of background factors to rater behaviour. The one significant exception is Pula and Huot's (1993) study of raters of English, rather than ESL, compositions, which interpreted its findings in light of Williamson's (1988) model of disciplinary enculturation. Certain aspects of raters' backgrounds, such as the range of their reading and writing activities, their professional training, and their work experience were seen as part of the process of socialization within an extended discourse community represented by English teachers. Other aspects, particularly membership in a "holistic scoring task group," represented socialization in the context of an immediate discourse community. Membership in the extended group was seen to explain the similarities in raters' scoring; membership (or its absence) in the narrower group the differences.

While Pula and Huot's study is a useful exercise in model-building, and employs a sound methodology in combining coded verbal protocols with confirmatory interviews, its conclusions are dangerously simplistic. Granted, for example, that the participants in their study reported a similar range of formative influences; however, such a statement is meaningless without an examination of the specifics of raters' reading experiences. Again, the idea that reading experiences led each rater to internalize a model for good writing may be sound, but the assertion that they all internalized the same model — implied by the statement that each
rater emphasized "content and organization as their primary criteria in scoring" (Pula & Huot 1993, p. 252) — is not. Like "reading experience," constructs such as "content" and "organization" have as many unique manifestations as there are raters; even if they could be authoritatively defined, it would be unbelievable, not to mention depressing, if all English composition teachers regarded them as the primary yardsticks for good writing. All of this leads to the conclusion that while discourse communities are a useful analytical concept, they are probably more numerous, and less easily defined, than assumed by Pula and Huot. Such a conclusion applies particularly to the culturally, linguistically and ideologically diverse community constituted by teachers of English as a Second or Foreign Language.

At the same time, as intended, the literature review did identify four potentially significant sources of variation in the backgrounds of raters of ESL compositions: cultural background, mother tongue, academic background, and professional experience. How I used this information to select participants for my study, and elicit data from them through interviews, is discussed in the next chapter concerning research design; to what extent my choice of these four background factors helped explain variability in the behaviour of the participants in the study is, in turn, addressed in the discussion in Chapter 8.
Chapter 3: Research Design

3.1 Participants

The four participants whose case studies I intended to use in answering the research questions I developed in Chapter 1 were selected from the ten who had taken part in the Cumming et al. (1998) study funded by Educational Testing Service (ETS) in the dual role of paid participant and researcher. They had all signed forms drawn up by ETS indicating their informed consent to the use of personal information about them, as well as their agreements to preserve the confidentiality of all information related to that project. In addition to providing the protocol and questionnaire data for the Cumming et al. (1998) study in their own homes or offices, they met regularly to discuss how the data generated by the project could be interpreted. Since they provided additional data in the present study, by granting an interview exploring the influence of their backgrounds on their rating behaviour during phase 1 of the Cumming et al. (1998) study, I asked them to sign additional forms (see Appendix A), approved by OISE/UT's ethical review committee, in which they gave their informed consent to my use of information relating to their background on condition that their identities be concealed and that they be free to withdraw from the study at any time. All the data that I collected from them for the present study came from interviews conducted in my office.

I chose the 4 participants because as a group they embodied the greatest variability along the following key dimensions (identified in my literature review presented in Chapter 2):

- Cultural background;
- Linguistic background, including not only raters' L1 but also the extent and nature of non-native speaking participants' use of English;
- Academic background and professional training, including details of degrees and certification obtained and institutions attended; and
- Professional experience, including teaching and assessment experience in general, and familiarity with specific rating scales in particular.

As Table 1 shows, in terms of linguistic and cultural background, Chris and Jo were native speakers (NS) from North America. Alex was from Asia, and Sam from Europe, but
while both were non-native speakers (NNS), they differed greatly in the extent to which they had used English in their native countries. In terms of academic background, Chris and Alex studied Natural Sciences while Jo and Sam obtained Arts degrees. While all the participants were "experienced," they represented different levels of experience. Sam had only limited training in teaching or assessing writing and no experience with assessment outside classroom situations, except as a candidate; Chris had received formal training in teaching and assessment, and had conducted placement tests; Alex and Jo had not only worked as raters for large-scale standardized assessments, but had been involved in test and (in Jo's case) rating scale development. Finally, Alex and Jo reported to me that they had been influenced by specific rating scales during the research, while Sam and Chris reported that they hadn't.

In addition to their differences the four also shared some attributes. They considered themselves to be either "competent" or "expert" assessors, even if they varied in level of experience. They all shared membership, as graduate students, in the same North American institution of higher education. In addition, they had all participated in Cumming et al.'s (1998) project concerning the behaviour of raters of ESL compositions, although given that the protocols which form the primary data base for this thesis were generated at the outset of the project, prior to any group discussions, it is unlikely that they had been influenced much in their behaviour by their membership in this particular, short-term discourse community.

Finally, as the use of androgynous pseudonyms shows, I decided not to take gender into account in explaining differences between the participants' behaviour, not because I did not consider gender as a potential influence on rater behaviour, but because I felt that the investigation of the effects of gender differences was beyond the scope of my study. My policy of masking participants' gender was furthered by my balancing act regarding pronoun usage, which was to refer to raters (both in the present study and as a generic category) in the masculine, and to writers in the feminine. I did this to protect the confidentiality of the individual participants.
### Table 1: Principal Characteristics of Participants

<table>
<thead>
<tr>
<th></th>
<th>Sam</th>
<th>Chris</th>
<th>Alex</th>
<th>Jo</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>30+</td>
<td>30+</td>
<td>40+</td>
<td>40+</td>
</tr>
<tr>
<td><strong>L1</strong></td>
<td>Language other than English</td>
<td>English</td>
<td>Language other than English</td>
<td>English</td>
</tr>
<tr>
<td><strong>Use of English in country of origin</strong></td>
<td>Limited to academic contexts</td>
<td>N/A</td>
<td>Extensive (academic and social contexts)</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>University Degrees Earned</strong></td>
<td>B.Ed., M.A.</td>
<td>B.Sc., M.A.</td>
<td>B.Sc., M.A., PhD in progress</td>
<td>B.A., M.Ed.</td>
</tr>
<tr>
<td><strong>Professional Certification</strong></td>
<td>Teachers' Certificate</td>
<td>TESL Certificate</td>
<td>Teachers' Certificate</td>
<td>TESL Certificate</td>
</tr>
<tr>
<td><strong>Years of ESL teaching experience in home country</strong></td>
<td>Primary &amp; Secondary level: 6 Tertiary level: 0</td>
<td>N/A</td>
<td>Primary &amp; Secondary level: 15 Tertiary level: 5</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Years of ESL teaching experience in Canada</strong></td>
<td>Private tutoring: 2 Tertiary level: 6</td>
<td>0 (at the time of the study)</td>
<td>Tertiary level: 12</td>
<td></td>
</tr>
<tr>
<td><strong>Years of experience in assessing L2 writing</strong></td>
<td>6 (in-class assessment only)</td>
<td>8 (in-class &amp; placement assessment)</td>
<td>12 (in-class &amp; standardized assessment)</td>
<td>19 (in-class &amp; standardized assessment)</td>
</tr>
<tr>
<td><strong>Additional assessment experience</strong></td>
<td>None</td>
<td>None</td>
<td>Experience with coordinating rater training</td>
<td>Experience with rating scale development &amp; rater training</td>
</tr>
<tr>
<td><strong>Self-assessment of competence</strong></td>
<td>Competent rater</td>
<td>Competent rater</td>
<td>Expert rater</td>
<td>Competent rater</td>
</tr>
<tr>
<td><strong>Influence of rating scales</strong></td>
<td>5-point holistic scale used in country of origin (without descriptors)</td>
<td>6-point rating scale used for placement at a North American university.</td>
<td>Jacobs et al. (1981); Hamp-Lyons (1991)</td>
<td>scoring guide for the Test of Written English (Educational Testing Service, 1989, p. 29)</td>
</tr>
</tbody>
</table>
3.2 Sources of data and instrumentation

Data for the study were integrated from three sources, described in detail below. Information on raters' processes of constructing scoring criteria was obtained from concurrent think-aloud protocols collected as part of a broader study on the behaviour of raters of ESL compositions (Cumming, et al., 1998), to which I have already referred and in which I participated. The influence of background factors on raters' strategies and judgements was in turn explored through a questionnaire, likewise designed during the Cumming et al. (1998) study, as well as through interviews which I designed and conducted independently.

Such multiple sources of evidence facilitated triangulation (Miles & Huberman, 1984, pp. 234-235) for my analysis of raters' construction of scoring criteria. The think-aloud protocols yielded qualitative information concerning cognitive processes which I regarded (following Ericsson & Simon 1993, p. 30) as unmediated by the participants' prior knowledge or purposes. This information I could relate to additional direct evidence — the scores participants assigned to compositions — as well as to evidence produced by participants' introspection in the course of their interviews.

Triangulation was more difficult to achieve in my analysis of the influence of background factors, however. My principal source of information on participants' backgrounds — interviews — and the principal checks on this source of information — raters' answers to a questionnaire and member-check verification (Miles & Huberman, 1984, p. 242) subsequent to the interviews — both relied on introspection. Although raters did occasionally make references to background factors in their concurrent think-aloud protocols, and these, unlike statements made in interviews or questionnaires, did constitute direct evidence of cognitive processes, only limited cross-checking of different sources of evidence was possible. However, before further detailing my methods of data analysis, the three principal instruments used in the study — concurrent think-aloud protocols, questionnaires, and interviews — must be outlined along with their relevance and limitations.
3.2.1 Concurrent think-aloud protocols

The concurrent think-aloud protocols which I used in the present study to examine the process raters of ESL compositions followed in constructing scoring criteria were originally provided during phase I of the Cumming et al. (1998) study of raters' decision-making behaviour, funded by Educational Testing Service (ETS). As part of that study, "10 experienced instructors of ESL/ EFL compositions — selected for the diversity and extent of their experiences internationally — each rated randomly sequenced samples of 60 TOEFL essays (selected by ETS staff to represent a range of 4 essay topics and 6 score points from a recent version of the test)" (Cumming, et al., 1998, p. 6) and furnished concurrent think-aloud protocols. I was, in turn, allowed to use data in the protocols on condition that I did not reproduce references to the TOEFL writing prompts or quotations from the essays written in response to these prompts.

The essays used in phase I of the Cumming et al. (1998) study, written in 30 minutes in response to a short essay prompt, were selected from a "recent administration" of the Test of English as a Foreign Language (TOEFL), a proficiency test widely used for screening NNS applicants to North American institutions of higher education. They formed a corpus of 144 compositions and each was given a unique serial number from #1 to #144, through which they could be referred to (both here and in the Cumming et al. (1998) study). Out of this corpus, participants were assigned sets of 60 compositions, selected using a table of random numbers. Neither the identity of the authors of the compositions, nor the scores assigned to the compositions during the actual administration of the test(s) were revealed. Instead, participants were asked to assess the compositions independently, and while they were instructed to assign a score using a 6-point scale, they were also instructed not to use the usual scoring rubric for TOEFL essays (Cumming, et al., 1998, p. 9). In addition to rating the compositions, participants were requested to produce concurrent think-aloud protocols "to document their decision-making behaviours" (Cumming, et al., 1998, p. 6) and were given one training
session in think-aloud procedures prior to their rating session. Following their rating sessions, participants were paired and asked to transcribe each others' protocols according to conventions developed during the project (Cumming, et al., 1998, p. 80-81). In the final stage of data collection, the transcriptions were verified by each participant.

Although, as discussed at the outset, the circumstances in which participants rendered their judgements — particularly the absence of a scoring rubric and the lack of practical consequences to participants' decisions regarding the quality of the compositions — were not reflective of typical practice in standardized tests, I, along with Cumming et al. (1998) felt that they would enable analysis to lay bare the processes involved in raters' construction of scoring criteria and to link the variability observed in raters' judgements and behaviours with variability in their backgrounds. Consequently, I chose the concurrent think-aloud protocols of four participants for analysis in my follow-up study, with the aim of maximizing rater variability in terms of mother tongue, academic background, teaching experience and experience with both large-scale writing assessment and specific rating scales, factors which I had identified as potentially significant in a study of raters' backgrounds based on a literature review (presented in Chapter 2, below). In order to conceal the identity of the four participants, I gave each a pseudonym that differed from those they had been assigned during the Cumming et al. (1998) study. Since all the protocols had already been transcribed by the time I began my study, I followed the original transcription conventions (Cumming et al., 1998, pp. 80-81 — reproduced here as Appendix B) in presenting excerpts from them.

The principal advantage of concurrent think-aloud protocols in the context of my study was that they provided evidence of cognitive processes that was not coloured by introspection (Ericsson & Simon 1993, p. 30). Although, as Green (1998, p. 4) rightly points out, concurrent think-aloud protocols do not report directly on cognitive processes, they do provide expressions of thoughts from which such processes may be inferred. Such evidence is particularly useful in construct validation, revealing in the present study, for example,
whether raters paid attention to a similar range of textual qualities in assessing compositions, and, by extension, whether they had the same construct in mind when assessing "writing quality." Further, such evidence can be analyzed both qualitatively and, once coded, quantitatively. An additional attraction is the sheer quantity of data yielded: in Cumming's (1990) study, for example, raters of ESL compositions averaged 23 to 32 coded comments for every composition they rated. It is, therefore, not surprising that think-aloud protocols, whether concurrent, retrospective, or both, have been widely used in efforts to investigate direct writing assessment (Connor & Carrell, 1993; Cumming, 1990; Huot, 1993; Pula & Huot, 1993; Vaughn, 1991; etc.).

This is not to say that the method is without drawbacks. One difficulty is that informants verbalize their thought processes to differing degrees. To overcome this problem, Green (1998, pp. 10-11) recommends providing consistent and explicit instructions to participants, a practice adopted by the study (Cumming, et al., 1998) which yielded the concurrent think-aloud protocols on which my own study depended. The second difficulty, that even in concurrent think-aloud protocols people cannot report on everything that goes on in their minds, is more serious, as it cannot be overcome by training. Interestingly, such a concern, not to mention an awareness of the distractions created by concurrent think-aloud protocols, was voiced by one of the participants in my study:

I may have not reported everything that was going on because while I was reading my decisions, I must say that most of the time that I was rating back home, I might have been using a mixed code in my mind when I am rating. I may, I may have, I don't know, I really don't know. I may have, I don't know but I may jump from my L1 to English in my mind. So when I was doing this, I am not sure, and I may not have reported everything that was going on because trying to do that would slow down the rating process very much.¹

Consequently, following the methods of Cumming et al. (1998, p. 9), I based my analysis of the data on "impressionistic interpretations of patterns and trends," supplemented by a quantitative component, involving frequency counts of coded behaviours, in a manner to be described in the next section.

¹ Note that to avoid confusion, excerpts from participants' interviews have been printed in Helvetica, and excerpts from protocols in Times New Roman.
One final danger was that concurrent think-aloud protocols could alter the behaviour of a participant who was not used to talking aloud while performing cognitive tasks. This was directly acknowledged by another of the participants in the present study who began the rating session by pulling out several short compositions that "were good candidates" for a score of 1, not for lack of extensive experience with rating compositions, but, as explained in the interview, simply to become comfortable with doing concurrent think-aloud protocols. In such a situation, the triangulation of data from multiple sources was particularly valuable, since erroneous interpretations of behaviour could be corrected.

3.2.2 The questionnaire
The questionnaire used in my study was also designed during the Cumming, et al. (1998, pp. 72-76) study and the ten participants in that study were requested to complete it immediately after their rating sessions in order to provide information on their personal and professional background. Although questionnaires are generally best suited to providing large data sets concerning categories, rather than nuances of individual attitudes and behaviour, they played an important role in the present analysis for two reasons. First, they facilitated my selection of participants for the study. Second, they provided a limited, supplementary check on participants' introspection as the principal source of information concerning background influences on their behaviour, helping to provide additional sources of data in the study.

3.2.3 Interviews
I designed, piloted and administered interviews following my initial analysis of the think-aloud protocols, independently of the project that had yielded the protocols and the questionnaire data (Cumming et al., 1998). My aim, besides providing a check on my analysis of raters' behaviours, was to explore raters' backgrounds, and to relate them to their rating behaviours. Interviews were suitable for the present study largely because they complemented
already available evidence: they allowed participants to explain, rather than merely report on, their behaviour and thought processes, something which they had been explicitly requested to avoid in the course of giving their think-aloud protocols. I designed the interviews to elicit information in a manner that was controlled, but also flexible in responding to the substance of each person's think-aloud data and, to counter the risk that I or the interviewees could bias the interview towards socially desirable interpretations of behaviour, I analyzed the internal consistency of responses and compared them with responses in the questionnaires to pinpoint answers that may have had such motives.

Part I of the interview was open-ended. Participants were asked to read excerpts of their concurrent verbal protocols involving 12 of the 60 compositions they had rated and commented on in the Cumming et al. (1998) study. The transcripts containing the excerpts were prepared so that all comments on a particular composition appeared on a separate page, in the order in which they had been uttered. In all, participants were asked to read 2,700 - 5,400 words, a task for which I had budgeted 45 minutes but which took, on average, 35 minutes.

The excerpts amounted to 21.4% to 25.8% of the original protocols provided by participants as measured by number of words; more importantly, they contained all comments made on 20% of the compositions in a corpus that a participant had assessed, made up of two compositions at each of six levels of proficiency in English as assigned by the participant during the rating session. The protocols were given to the participants to comment on based on the scores assigned to the compositions they related to: protocols involving compositions with a score of 1 were given first; protocols of compositions scored 6 were given last. The purpose of this arrangement was to remind participants of the textual features they attended to at each successive level of proficiency. Whenever possible, the first and last essays rated at each level of proficiency were chosen, to reduce subjective criteria for selection which may influence raters' own recall of the rating sessions. In practice, this principle of selection was modified for the following reasons: a) protocols concerning compositions #14, #62, and #128,
the only ones which were rated by all four participants in my study, were included in each interview because they afforded an opportunity to make direct comparisons between the way different participants reacted to the same compositions; b) protocols where raters verbalized a scoring scheme by listing the textual features they associated with different levels of proficiency were at times included even if the compositions raters were commenting on were not the first or last of their respective levels of proficiency to be rated.

The purpose of this part of the interview was, first of all, to get participants to recall their rating sessions in general, and their scoring criteria and rating strategies, in particular. This was considered necessary in view of the amount of time (between 6 and 7 months, depending on the participant) that elapsed between the concurrent think-aloud protocols and the interviews. In addition, by asking participants to comment on their protocols specifically with respect to their backgrounds, I intended to raise their awareness of the range of background factors potentially pertinent to their behaviour, in order to prepare for the specific questions subsequently posed in Part II of the interview.

In terms of procedures, once participants completed their reading of the protocols, I asked them to comment on the protocols involving each composition in turn, in the order already described (those ranked lowest, first, and those ranked highest, last); then I asked them for general impressions concerning their rating session. The only instruction accompanying the protocols, printed on each page of the transcripts, was that they should "comment on the protocols with reference to your background as a learner, teacher and/or assessor of English, ESL, or any other language." Although a list of background factors, having been identified in my literature review as potentially significant, was kept at hand to prompt participants, I referred to them only if it was absolutely necessary to keep the flow of information going. Otherwise, I wanted participants to comment freely, partly to allow factors other than those I had already identified as significant to emerge, and partly because, as Pula and Huot (1993, p. 241) noted in their related study of raters' backgrounds, "data volunteered by infor-
ments may be more valid than data generated by a specific, and perhaps closed-ended, prompt." Including the time spent by participants reading their protocols, I had budgeted 90 minutes for conducting Part I; the time actually spent on this part eventually ranged between 80 and 110 minutes.

Part II of each interview followed immediately after part I, and consisted of a question-and-answer session based on my analysis of participants' behaviour. Its purpose was two-fold: to get raters to provide comparable data by relating key dimensions of their behaviour to their backgrounds and to provide an initial "member check" on my analysis of their rating behaviour. The framework for constructing this part of the interview was provided by my research questions which had highlighted four key stages in the process of direct writing assessment, thus affording four broad areas of comparison between raters: a) their attitudes concerning the relationship of performance to proficiency, b) their interpretations of the role of writing prompts in direct assessment, c) their reading strategies (concerning both individual compositions and a corpus of compositions); and d) their methods of constructing specific rating criteria. In analyzing the concurrent verbal protocols, I identified key dimensions of variability at each of the four stages of the rating process whenever a comparison of two raters showed differences in strategies and judgements which were subsequently identifiable in the other two participants. Such a procedure followed the precepts of "grounded theory" (Glaser & Strauss, 1967) in that differences between raters emerged from the data without being hypothesized prior to analysis. Having identified these dimensions of variability, I could design interviews that provided consistency, in that the questions to each of the four participants concerned the same dimensions of variability, and flexibility, in that the questions were worded to reflect specific differences in behaviour along these key dimensions.

To illustrate my methods, one example of the analytical procedures I followed began with my observation, based on my analysis of the protocols, that a key source of variability in how participants estimated proficiency from performance lay in their attitudes to situational
factors: two rated the compositions at face value and two took situational factors (particularly, time pressure and topic effects) into account. In the interviews I then asked each participant to comment on their specific attitudes to the role of situational factors in performance on a composition test in terms of their backgrounds. Designed in such a way, the interviews ensured consistency by exploring an identical range of issues across the four participants (here concerning their attitudes to the potential constraints exerted by time pressure and inappropriate topics), but remained flexible by allowing me to use specific observations concerning participants' behaviour in phrasing each question. Participants' responses could also be compared with their statements concerning the establishment or revision of rating criteria in their protocols, as well as against statements they made in the questionnaire concerning background and attitudes.

I expected Part II of the interview to take about 90 minutes to conduct and this estimate proved to be accurate. At the conclusion of each interview, both part I and part II were transcribed with all references to participants' ethnicity, as well as to educational institutions or professional organizations they had been, or were being, associated with, omitted. The transcriptions followed the conventions already established by the Cumming et al. (1998) study for concurrent think-aloud protocols (and reproduced in Appendix B here).

Finally, at the conclusion of my analyses, participants were presented with their own case studies (as well as parts of the thesis that would present the broader context for the analysis) for comments that were designed as a follow-up "member check" on my interpretation of the relationship between their backgrounds and their behaviours. This follow-up interview, 5 to 7 months subsequent to the first interview, was entirely open-ended; I simply asked participants whether they felt that my interpretations of their behaviours were accurate, and if they wished to raise any specific concerns they had about omissions and inaccuracies. I was particularly interested in their opinions regarding my use of extensive quotations from both their concurrent think-aloud protocols and their interview transcripts.
3.3 Methods of analysis

As discussed in 3.2.3, my analysis of the concurrent verbal protocols and my construction of the interviews followed the precepts of "grounded theory" (Glaser & Strauss, 1967) in that differences between raters emerged from the data without being hypothesized prior to analysis. However, in addition to qualitative methods in analyzing the protocol data, I also incorporated a quantitative component aimed at highlighting the textual features raters attended to during various stages of their rating session. This relied on frequency counts of behaviours coded according to a scheme developed in earlier analyses of the same think-aloud data (Cumming, et al., 1998, p. 12-15). The coding scheme (reproduced in Table 2; see also Appendix C for a sample coded protocol) was validated by its "precision and relevance" in describing raters' behaviours in assessing a wide range of essays responding to 6 TOEFL prototype tasks in the Cumming, et al. (1998, pp. 45-48, 83-89) study. It is based on two key distinctions. The first one distinguishes interpretation strategies and judgement strategies and the second, cross-cutting the first, establishes three foci for such strategies: self-monitoring, rhetorical-ideational, and linguistic. Although alternative taxonomies (e.g., Pula & Huot, 1993; Vaughn, 1991) could have been chosen, the scheme employed here has not only been refined through extensive use, but has also proven to yield reliable codings of protocols (Cumming, et al., 1998, p. 32). In adapting it to the present study I added two new codes: "ISO" ("read essay prompt") and "JRO" ("consider writer's use and/or understanding of the prompt"), to reflect my specific interest in raters' attitudes to writing prompts; these new codes complement code "IS1," which in the original scheme covered the full spectrum of raters' comments concerning writing prompts but is restricted in the present study to behaviour labelled as "judge level of difficulty of (or level of challenge posed by) essay prompt."

During phase II of the Cumming et al. (1998) study, I had been one of two participants who had used the coding scheme (Table 2) to independently code, with an average inter-coder
Table 2: Coding Scheme of Decisions Made While Rating ESL Compositions (based on Cumming, et al., 1998, p. 48)

<table>
<thead>
<tr>
<th>Self-monitoring focus</th>
<th>Rhetorical and Ideational focus</th>
<th>Language focus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interpretation strategies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS0 read essay prompt</td>
<td>IR1 interpret ambiguous or unclear phrases</td>
<td>IL1 observe layout</td>
</tr>
<tr>
<td>IS1 judge level of difficulty of (or level of challenge posed by) essay prompt</td>
<td>IR2 discern rhetorical structure</td>
<td>IL2 classify errors into types</td>
</tr>
<tr>
<td>IS2 Read/reread composition</td>
<td>IR3 summarize ideas or propositions</td>
<td>IL3 &quot;edit&quot; phrases for interpretation</td>
</tr>
<tr>
<td>IS3 envision personal situation of writer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS4 scan composition</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Judgement strategies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JS1 decide on macro-strategy for reading and rating (either for a paper or for a corpus)</td>
<td>JR0 consider writer's use and/or understanding of prompt</td>
<td>JL1 assess quantity of total written production</td>
</tr>
<tr>
<td>JS2 consider own personal response or biases</td>
<td>JR1 assess reasoning, logic, or topic development</td>
<td>JL2 assess comprehensibility</td>
</tr>
<tr>
<td>JS3 define and/or revise own criteria</td>
<td>JR2 assess task completion</td>
<td>JL3 consider gravity of errors</td>
</tr>
<tr>
<td>JS4 compare with other compositions or &quot;anchors&quot;</td>
<td>JR3 assess relevance</td>
<td>JL4 consider error frequency</td>
</tr>
<tr>
<td>JS5 summarize, distinguish, and tally judgements collectively</td>
<td>JR4 assess coherence</td>
<td>JL5 assess fluency</td>
</tr>
<tr>
<td>JS6 articulate general impression</td>
<td>JR5 assess interest, originality, creativity, sophistication</td>
<td>JL6 consider lexis</td>
</tr>
<tr>
<td>JS7 articulate or revise scoring decision</td>
<td>JR6 identify redundancies</td>
<td>JL7 consider syntax and morphology</td>
</tr>
<tr>
<td></td>
<td>JR7 assess text organization</td>
<td>JL8 consider spelling and punctuation</td>
</tr>
<tr>
<td></td>
<td>JR8 assess style, register, or genre</td>
<td>JL9 rate language overall</td>
</tr>
</tbody>
</table>
agreement of 84% (Cumming et al., 1998, p. 32), a sample of transcribed protocols which included participants' assessments of 20 of the 144 compositions at various levels of proficiency. For the present study, I subsequently coded the full transcripts of the rating sessions of the 4 raters whom I had requested to participate. Then, having broken down each participant's rating session into distinct stages (described in the case studies in Chapters 4-7), I tabulated the total number of coded comments with self-monitoring, rhetorical-ideational, and linguistic foci for interpretation strategies as well as for judgement strategies for each phase. Finally, to allow comparisons between participants, I converted the data to show the proportional frequencies of comments with self-monitoring, rhetorical-ideational, and linguistic foci for both interpretation strategies and judgement strategies for the different phases of participants' rating sessions, and for ease of display I represented them in pie-charts (Figures 1 to 26). The data thus obtained supplemented my qualitative analyses of participants' comments, and helped me to determine the textual qualities that they paid attention to while reading and assessing the compositions.

Unlike my analysis of the concurrent verbal protocols, however, my analysis of the interviews was strictly qualitative and involved three cycles of interpretive comparisons. Having identified key contrasts during each of the four stages of the rating process in the protocol data, I first looked for confirmation from the participants that such contrasts were indeed present. Subsequently, I sought to identify aspects of their background that could explain the contrasts by triangulating data from the interviews, the verbal protocols and the questionnaires. I documented these contrasts, and the factors likely responsible for them, in each case study in Chapters 4 to 7. Finally, by making comparisons across the participants, I was able to highlight not only aspects of raters' backgrounds which influenced each stage in their assessment of ESL compositions, but also aspects of their backgrounds where variability was most clearly related to variability in their judgements and behaviour. The results of this final cycle of interpretive comparisons is presented in the discussion in Chapter 8.
3.4 Construction of the case studies

The 4 case studies played a central role in my research design and in order to answer my research questions I had to ensure that they were comparable. Consequently, I structured Chapters 4-7 in such a way that the case studies not only addressed the 4 research questions in the same sequence, but addressed the same sources of variability, in the same order, using identical sources of data and identical methods of analysis. Therefore, to conclude the discussion here, I have summarized the structure of the 4 case studies in Table 3 by listing the key sources of variability pertinent to each research question, the data which contained information on these sources of variability, and the methods of analysis I used to demonstrate that such variability actually existed. Such a table provides not only a convenient summary of my research methods but also a convenient road map for the chapters to follow.
Table 3: The structure of the case studies

<table>
<thead>
<tr>
<th>Sources of variability</th>
<th>Sources of data</th>
<th>Methods of analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Research Question A</strong></td>
<td>- statements concerning the nature of language proficiency and language learning (in the protocols or the interviews)</td>
<td>- interpretive analysis and comparisons between participants</td>
</tr>
<tr>
<td>- views on language proficiency and language learning</td>
<td>- statements relating performances on composition tests to writing proficiency (in the questionnaire or the interviews) with reference to a developmental trajectory for language learners</td>
<td>- interpretive analysis and comparisons between participants</td>
</tr>
<tr>
<td>- views relating performance to proficiency</td>
<td>- scores assigned to unfinished essays (in the protocols) and comments on such scores (in the interviews)</td>
<td>- interpretive analysis and comparisons between participants</td>
</tr>
<tr>
<td>- attitudes to situational factors affecting performance</td>
<td>- comments (in the protocols or the interviews) inferring a writer's potential from her performance</td>
<td>- interpretive analysis and comparisons between participants</td>
</tr>
<tr>
<td><strong>Research Question B</strong></td>
<td>- participants' analysis of the cognitive demands set by the prompts (in the protocols or the interviews)</td>
<td>- interpretive analysis and comparisons between participants</td>
</tr>
<tr>
<td>- views on the role of writing prompts</td>
<td>- frequency counts of references to task fulfilment (JR2) or relevance (JR3) (in the protocols)</td>
<td>- interpretive analysis and comparisons between participants</td>
</tr>
<tr>
<td>- views on biases in writing prompts</td>
<td>- frequency counts of references (in the protocols) to test takers' ethno-linguistic or socio-cultural backgrounds</td>
<td>- interpretive analysis and comparisons between participants</td>
</tr>
<tr>
<td>- performance expectations</td>
<td>- statements (in the protocols or the interviews) concerning biases in the writing prompts</td>
<td>- interpretive analysis and comparisons between participants</td>
</tr>
<tr>
<td>- assumptions concerning test takers (in the interviews)</td>
<td>- assumptions concerning test use (in the interviews)</td>
<td>- interpretive analysis and comparisons between participants</td>
</tr>
</tbody>
</table>
| Research Question C | - sequences of coded behaviours within participants' protocols dealing with individual compositions  
- participants' comments (in the interviews) concerning their reading of individual compositions 
- sequences of coded behaviours within participants' rating sessions as a whole  
- participants' comments (in the protocols and the interviews) concerning their strategies for reading a corpus of compositions | - interpretive comparisons between protocols provided by the same participant  
- interpretive comparisons between protocols provided by the same participant  
- interpretive analysis and comparisons between participants |
|---|---|---|
| Research Question D | - interpretive and judgemental comments concerning textual quality in participants' protocols  
- statements concerning relative importance of linguistic and rhetorical-ideational features (in the interviews)  
- interpretive and judgemental comments concerning textual quality in participants' protocols  
- statements establishing criteria for specific scores (in the protocols)  
- participants' views concerning a developmental trajectory for language learning (in the protocols or the interviews)  
- proficiency levels established by participants for the lowest and highest points in their rating scale (in the protocols or the interviews)  
- pass/fail thresholds established by participants (in the protocols or the interviews)  
- specific performance expectations (as discussed under Research Question B) | - proportional frequencies of comments with a linguistic focus and comments with a rhetorical-ideational focus (calculated separately for interpretive and judgemental behaviours, and for each phase of a participant's rating session — see Figures 1 to 26)  
- interpretive analysis and comparisons between participants  
- comparisons of proportional frequencies of comments concerning specific textual qualities  
- interpretive analysis and comparisons between participants  
- interpretive analysis of participants' comments linking proficiency levels to the mastery of specific competencies, and comparisons between participants  
- interpretive analysis and comparisons between participants  
- interpretive analysis and comparisons between participants  
- interpretive analysis and comparisons between participants |
Chapter 4: Sam

At the time of the study Sam was a graduate student in his 30s. He had come from Eastern Europe to Canada two years previously. Although a certified teacher in his native country for 6 years, his experience in assessing English was limited to classroom evaluation, his only encounter with standardized assessment having been in the role of a test-taker. I chose him for my study because in spite of his relatively limited assessment experience he characterized himself as a competent rater, and this self-perception was reflected in the consistency and confidence with which he assessed compositions and articulated his impressions; I felt that finding the roots of such consistency and confidence in Sam's personal and professional experiences held considerable interest for a study that aimed at explaining raters' behaviours in terms of their background.

4.1 Research question A: In what principled way did the raters of ESL compositions participating in the study relate the performance embodied in a written text to a writer's level of proficiency, and how did background factors influence this process?

Sam's protocols contain 33 statements, representing 2.7% of all his coded statements, that were coded as "envisioning the writers of compositions" (IS3 — see Table 2). Several of these comments were intended to specifically relate a writer's observed performance to her level of proficiency, usually by using the mastery (or absence of mastery) of specific competencies as a guide:

[ESSAY #56] has so many words, just words, words, and words and I think it shows that the person thinks he or she knows a lot of English but actually doesn't have [...] a good command of the language [...] doesn't have good knowledge of how essays should be structured and framed.

Such behaviour recalls the view (Prior 1995, p. 53) that readers' judgements of graduate students' compositions are at least partly based on inferences about students' effort, thought, and knowledge; in the context of my study, it suggests that raters try to estimate a writer's proficiency from her performance, in this case an essay written under specific conditions. Indeed, in the protocols Sam went further by estimating not only a writer's proficiency, but, in some cases, also her potential, as with essay #113 (see also the protocol for essay #7):
I think it's possible that this person can quickly learn how to write a well-structured and well-organized essay. So, since this ought to be a placement test, I think it's more fair to give the person a 4. I guess, my grading is directed towards the person, not towards what the actual essay is. For the essay I'd give a 3, but for the person, for the sake of the person I'm giving a 4.

What further distinguished Sam was that he took a broad view of language proficiency, at times labelling essays as "intelligent" (#26), or "unintelligent," (#10, #62), and writers as "immature" (#113), "confident" (#40) or "not serious" (#14). Finally, in assessing compositions, he weighed the potentially negative impact of situational factors, even when these were impossible to pin-point. For example, even though essay #126 broke off after two sentences for reasons he was unable to identify, failing thus to meet even minimum requirements for task fulfilment, he rated it a 2 because the sentences themselves were well formed.

Since a rater's interpretation of the meaning of performance plays a vital role in the establishment of scoring criteria, it is worth exploring Sam's perception of how he came to hold his views. Based on my interview with him, the point of departure is to be found in his cultural background, in this case channelled through his teaching experiences in his native country, which had led him to place all learners on a developmental trajectory, and to define his rating task as one of identifying a writer's position on that trajectory. Significantly, this was the very first subject Sam raised, spontaneously, during his interview:

The educational system that I have been through has conditioned me so much. [...] Like, on the one hand you have the knowledge that the student is supposed to acquire and then you have the developmental aspects of the students. So as a teacher I would always look at those two in some sort of a separate way [...] knowledge is more like "this should be the end result in terms of some quantity or quality of knowledge" and then there is the person who is supposed to be developing in terms of skills, or attitudes or that type of thing. [...] I always look not only at what is being produced but also at what level of development those students are at that point, when they wrote that. So that, I guess, affects my assessment also; not only the actual production, the product that I have to assess, but I am somehow connecting it to the personal development that I think I can discern from their writing.

This view was reinforced by Sam's contacts as a teacher with students in a wide range of situations in and beyond the classroom, which had enabled him to construct a rounded picture of writers' development and personalities. These profiles could then be related to writers' performances on tests, facilitating the estimation of their potential from their current position on a developmental trajectory. Later in his career, working as an editor, Sam likewise matched writers' output with their abilities and personality, thus reinforcing his teaching experiences.
To complement his practice of constructing developmental trajectories and writer profiles, Sam's experiences as a learner and test-taker led him to conclude that every test, writing tests included, measured intellectual development in general as well as domain-specific knowledge:

I don't think the TOEFL test is only a language test [...] the assessment of intelligence always plays a role, even if we don't realize that, even if it's subconscious [...]. Because, in writing you're, you're exercising your intelligence. Writing is not simply language, [...] it's intelligence also, and assessing the written language product is assessing intelligence. [...] I am not saying that this is just an intelligence test. But the writing reflects one's intelligence, or, not intelligence but intellectual development, that's why I use the word development, because that, it shows the level which you are at now.

Such a broad definition of the scope of language tests had significant repercussions for Sam's construction of rating criteria (see 4.4), since "intellectual development" could be signalled by a much wider range of competencies than mere "writing proficiency." Conversely, Sam could declare that certain aspects of a writer's performance, such as adherence to textbook prescriptions, were irrelevant to the estimation of proficiency on the grounds that they had no bearing on estimating intelligence, but were the products of situational factors:

I have seen how people perform differently under stress, and I take that into consideration, and it comes from my experience as a teacher, because I think you can tell from the kind of writing that they do. You can tell if it is logical, if the language is grammatically correct, but, for example, the arguments are not very well developed. Then I can say that person probably has problems concentrating under stress, or had a problem with the topic, so I would consider those problems to be independent of what the student is able to do. So those kinds of problems, I think I can say, are not connected, or are not indicative of the abilities of the person. They are more situational.

To sum up, in the process of scoring, Sam not only related performance to proficiency by placing writers on a developmental trajectory, but also interpreted proficiency in a broad manner and took a writer's potential and, to a lesser extent, situational factors, into account. His interviews revealed that his attitudes were influenced by a wide range of personal and professional experiences. Although he relied most heavily on his mutually reinforcing experiences as teacher and student in the same educational system, he was also able to compensate for his lack of experience in standardized assessment by calling on his experiences as a test-taker, and on his observations regarding classroom assessment. Even his observations as an editor could be called upon. At the same time, having had to draw on a wide range of practical
experiences did not prevent him from developing a consistent policy of relating the quality of an individual composition to his perceptions of its writer's level of proficiency.

4.2 Research question B: How did raters of ESL compositions interpret the role of writing prompts in performance-based assessment, and how did background factors influence this process?

At the outset of his rating session, Sam analyzed each of the four writing prompts that writers in his corpus of compositions responded to, focusing particularly on the potential of each for eliciting a complex response. He found Topic A "very limiting"; Topic B was "better […] because it [ALLOWED] people to give interesting arguments"; Topic C was simple but, like Topic B, allowed for good argumentation and was "relevant to students"; Topic D was the least satisfactory because it forced writers to choose between two simple alternatives. Then, in subsequent protocols, he assessed the level of challenge posed by the writing prompts a further 8 times. Nor was he averse to criticizing prompts that placed unrealistic expectations on disadvantaged candidates, as his comments concerning essay #24 show. He also showed his sensitivity to writers' personal backgrounds by making inferences concerning their gender (#20, #21) and socio-economic background (#24), and by noting all references made by writers themselves to their backgrounds (#18, #40, #128). In one case (essay #56) Sam even speculated on the ethno-linguistic background of a writer in spite of the absence of any direct indicators.

In spite of such sensitivities, however, Sam clearly expected writers to cope with the challenge posed by inappropriate writing prompts:

I acknowledge the fact that students or test-takers are […] sometimes put in a position to answer stupid questions or they cannot relate to the topic because the topic is put in very inappropriate ways. […] So I'll take [WRITING PROMPTS] into consideration when I approach their writing (BUT) I also rate their ability to cope with the task.

The apparent contradiction can be resolved by examining the interplay of the background factors shaping Sam's attitudes. To begin with, in his interview he traced his interest in the ethno-linguistic background of writers to his early teaching experience in his native country:

I was for two years a teacher [IN A CLASS OF ETHNIC MINORITIES] and that was a huge shock to me, when I encountered that because […] I wasn't prepared to meet that type of environment
in [NATIVE COUNTRY], but during the period I did grow to understand cultural differences, to appreciate them. So that's something personal, just, just my personal experience.

This interest was taken a step further when Sam's exposure to foreign language texts (or their translations) led him to actually associate certain types of performance with certain ethno-linguistic backgrounds. At the same time, in his interview he ascribed his interest in the ethnicity of writers to a general, academic interest in anthropology, so it would be misleading to interpret his interest to a desire to compensate for systemic inequalities in this regard:

My interest in anthropology, my openness [IS WHAT LED ME TO NOTICE ETHNICITY] ... It's more of a notice than a criterion. So it's just something that is interesting to me, cultural differences that are reflected in the writing are interesting to me. So it's just a personal interest.

In addition, although Sam did not report this, I inferred that a combination of his extensive teaching experience and his limited assessment experience had conditioned him to view writers as complex individuals, and in the anonymous setting of a standardized test his only means of reconstructing writers' personalities lay in speculating about their backgrounds as reflected in their writings.

By contrast, Sam's interest in socio-economic background was explicitly tied to his feeling, arising from personal as well as teaching experience, that people can be seriously disadvantaged by their circumstances:

I think in my personal life experience [...] and also as a teacher I have seen people coming from different socio-cultural backgrounds and, and have seen the struggles that some of those people have had to go through in order to perform at the same level as other people from a much favourable social background. So I think we should, I think it is fair to allow for consideration, you know, to just to consider how much effort this person had to put in, and how easy was it for the other one?

That he nevertheless gave limited leeway to learners can be attributed, first of all, to his broad conceptualization of proficiency as "intellectual development," which meant that in his view writers handicapped by a particular topic had a range of options to overcome adversity, as he himself had done:

Often, as a student, I was supposed to write on topics that were ideologically loaded and I was not subscribing to that ideology, so that was a kind of opposition, just acknowledging the stupidity of the ideology put into the exam questions. [...] I was irritated and that stimulated my critical approach to things but at the same time I think that critical approach earned me a higher grades.
Conversely, if assessors were inferring not only linguistic proficiency but intellectual development, they had a wide range of criteria to aid their assessment, rendering overt compensation unnecessary. As evident from the previous quote, critical thinking was one key sign of "intellectual development." Another was the ability to use sophisticated linguistic devices, such as the conditional mode to address topics beyond a writer's personal experience; this was Sam's specific recommendation for coping with Topic "A," which he had found highly discriminatory.

In addition, Sam also made it clear in the interviews that he regarded the TOEFL not only as a test of intelligence, but, more specifically, as a measure of readiness for university, and saw the ability to strategically compensate for handicaps as a sign of the kind of intellectual maturity that those who write the TOEFL test, i.e., potential university students, must possess, if they are to succeed:

*My experiences as a teacher have taught me that students with good academic achievements are able to cope with tasks [...] usually show ability to understand what a task is about. And people who cannot understand that, usually that's a measure of intelligence and how schooled you are. If you are schooled then you are usually taught how to do that and to be a university student, you have to have that.*

Finally, Sam's broad definition of "proficiency" was complemented by a relaxed approach to writing prompts. Judgements concerning "relevance" were rendered only 10 times throughout his protocols, and judgements concerning "task fulfilment" only 13 times, representing 0.8% and 1.1% of all coded behaviours. Although he didn't state it explicitly in either the protocols, or in the interview, it is possible to conclude that Sam viewed the writing prompts as vehicles for eliciting a performance. This would explain not only his relative lack of interest in task fulfilment, but also his attitude, reflected in his initial analysis of writing prompts, that the more challenging a prompt, and the more complex the response that it permitted, the better. This attitude allowed him not only to overlook occasional misinterpretations of the prompt on the part of writers, but also to refuse to compensate writers he himself may have identified as disadvantaged: since prompts were merely vehicles for eliciting language perfor-
mance, any disadvantages that a writer may have suffered from should have been surmounted through appropriate strategies and, above all, linguistic competence.

4.3 Research question C: What reading strategies did raters of ESL compositions establish (to deal with both individual compositions and a corpus of compositions) and how did background factors influence these?

Sam's strategies in working through the corpus of compositions, and in rating individual essays were consistent, and were rooted in his practice of taking only a single pass through each essay and rendering a score immediately after this reading, exemplifying the "read-through" approach as defined by Milanovic, Saville and Shuhong (1995). His rating session broke down into three stages:

1) identify the composition and comment on its length and layout;
2) read the composition (usually silently), and provide a running commentary attending to both language and rhetorical-organizational features;
3) provide a final summation of judgements (with linguistic and rhetorical-organizational features weighed, again, fairly evenly) along with a score.

After a brief discussion of the essay prompts at the beginning of the session, and a brief scanning of the essays for length and layout, Sam followed this sequence of actions for all compositions from the beginning, with only minor variations (chief of which was that Sam resorted to editing and classifying errors more frequently as the rating session progressed). To illustrate the stability of this pattern, one need only compare the structure of Sam's first and penultimate protocols (with the beginning of each of the three stages marked):

<table>
<thead>
<tr>
<th>Essay #14</th>
<th>Essay #132</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(stage 1)</strong> Let's start with the first one, which is obviously not a good one, because of it's appearance, and it's handwritten.</td>
<td><strong>(stage 1)</strong> Next one, 132, a long one, typed, over half a page.</td>
</tr>
<tr>
<td><strong>(stage 2)</strong> I can read it now. So, reading, reading essay number 14. Reading silently. [...] I'm reading silently and [the answer] already tells me that [the writer is] either a teaser, or not a very serious person. O.K. ... I'm reading ... I'm reading and</td>
<td><strong>(stage 2)</strong> I'm reading it. Good beginning. Oh, [QUOTE] verb [problem] there ... Uh, OK some style here. I'm reading on, uh, there are problems with plural [QUOTE]. Uh, run-on sentence. Very bad sentence structure ... Uh, poor word choice ... Uh, poor sentence structure. I'm reading on. Again poor sentence structure, wrong verb tense, Bad word choice. I'm reading on [QUOTE]. Wrong preposition. Uh, bad phrase [QUOTE]. Some ambiguous word choices, I, I, I don't really know what</td>
</tr>
</tbody>
</table>
(stage 3) I'm finished reading, and I'm rereading some of the sentences actually. And, it's poor language, uh, no structure, broken sentences, and uh, poor vocabulary. And to me it's a 2.

(4 stage 3) So essay 132, a long essay, many words there, many wrong words, uh, poor phrasing, uh, bad sentence structure. There is overall essay structure, an attempt to write a good introduction, which is quite unsuccessful because of poor language. Also the body where arguments are developed, somewhat developed, but not convincing, and uh, a conclusion. So there are some traces of framing. Obviously this has been learned, that there should be framing, there should be this overall structure, but language wise poor. So 132 is a 3.

In the interview, Sam's use of the read-through strategy, the simplest method in Milanovic, Saville and Shuhong's (1995) taxonomy, emerged as the product of his teaching experience: lack of time for feedback in his teaching induced him to read compositions submitted by his students only once, and confidence in his own assessments arising out of years of teaching and writing experience obviated for him the need for confirmation even in a standardized test. As for Sam's practice of giving a running commentary, the other striking feature of his reading strategy, he saw it rooted in his need to be engaged with the written text as well as its author, a need which he explained as part of his personality:

I am not the type of person ... who can listen to what a person is saying without any judgement. [Some raters] just listen or they would just read, and then, they'd start thinking about what their reaction to that is. I cannot do that. My reaction is immediate. ... I cannot distance myself when I am reading, I cannot stop myself from judging. And I think that's personality.

4.4 Research question D: By what process did raters of ESL compositions construct specific rating criteria, and how did background factors influence this process?

Since Sam "defined or revised rating criteria" (JS3) 36 times in his protocols, I was able to reconstruct his reasons for assigning specific scores, and then explore them in his interview. Unlike some of the other raters, he neither used the qualities of the corpus at his disposal in constructing a rating scale, nor referred to any external norms in his protocols, preferring to
judge the compositions by absolute criteria; the strategy labelled "compare with other compositions or 'anchors'" (JS4) occurred but once in his protocols.

Proportional frequencies of his coded behaviours, presented in pie-charts (Figures 1 and 2), suggest that, overall, Sam judged language more frequently than rhetoric and ideas: his comments with a language focus accounted for 36.2% of all interpretive and 45.0% of all judgemental behaviours, while his comments with a rhetorical-ideational focus accounted for 6.4% of all interpretive and 32.8% of all judgemental behaviours. Such a pattern persists even if we break down Sam's protocols into the three stages previously identified. In the initial stage of his rating session Sam's comments were limited to length and appearance. However, while Sam often praised long, and typed or neatly written essays, he made no references to appearance and length in his final summation, when he actually awarded the score, in spite of the fact that he at times speculated on the score that a paper might receive based on its length.

During the running commentary on each composition, making up the second phase, interpretive strategies predominated, accounting for 65% of all coded behaviours. In addition, Sam's comments, particularly judgemental comments, focused heavily on language, as shown in the relative frequencies of coded statements in Figures 3 to 4. This trend intensified as he proceeded through the corpus, so that by the end of the session he had shifted to almost a proofreading mode. More significantly, there was a gradual shift in Sam's focus from language to ideas and rhetoric as papers were assigned successively higher grades, with the exception of compositions he rated as 6, where he appears to have been so impressed with language that his interest in rhetorical structure was put aside. Such findings, plus the fact that a concern with style and originality emerged only in papers Sam scored 3 or higher, recall Pollitt and Murray's (1995) sample of composition raters, who also shifted from commenting on language to commenting on argumentation, and, eventually, style as compositions of progressively higher levels were encountered.

It was during the final summation, representing phase III of his reading of individual compositions, that Sam invariably assigned a score to each composition. At this stage, judge-
Figure 1: Proportional frequencies of interpretive strategies during Sam’s rating session.

Figure 2: Proportional frequencies of judgemental strategies during Sam’s rating session.
Figure 3: Proportional frequencies of interpretive strategies during phase II of Sam's rating session.

Figure 4: Proportional frequencies of judgemental strategies during phase II of Sam's rating session.
ment strategies predominated over interpretive strategies, accounting for 88% of all coded behaviours, as would be expected. While, as shown in Figure 5, interpretive comments continued to focus on language, judgemental comments, shown in Figure 6, were fairly evenly balanced between language and ideas and rhetoric, and Sam's focus once again shifted gradually from language to rhetoric and ideas as his scores increased. The critical variables Sam considered were few in number, although none can explain a score given in isolation: these included: "reasoning, logic, or topic development," "text organization," "length," "lexis," "syntax and morphology," and "overall language ability."

On the whole, then, Sam took a balanced view of proficiency, and weighed both the quality and organization of ideas and the mastery of linguistic conventions to produce a score. On the evidence of his protocols, an essay that he rated as a 1 exhibited, in his view, no organization and either no development of ideas, or, indeed, no ideas at all, besides lacking language control even at the sentence level. Essays he rated as a 2 showed evidence of either rudimentary language control (e.g., #11, #62) or rudimentary organization (e.g., #15, #137). In essays he rated a 3 there emerged argumentation; in addition Sam concluded that most, but generally not all, of the requisite structural elements of an essay were present and paragraphing was mastered. On the other hand, Sam found that the language continued to lack proficiency (e.g., #117, #37). Level 4, while still considered a "fail," showed improved control of language and/or evidence of original thought (e.g., #98).

Sam's verbal reports suggested that the boundary between 4 and 5 was relatively permeable, perhaps because one was deemed a fail and the other a pass, and the consequences of failing the test were considered to be sufficiently serious to influence Sam's judgement, as the following comment shows:

I think that essay #7 deserves a 4, and not a 5. I was wondering if it should be a 4 or a 5, but, uh, a 4. If I knew that the 4 would not be a pass grade for this person to be accepted to a university, I would give a 5.
Figure 5: Proportional frequencies of interpretive strategies during phase III of Sam's rating session.

Figure 6: Proportional frequencies of judgemental strategies during phase III of Sam's rating session.
On the whole, however, Sam felt that to get a 5 a writer had to develop her arguments fully, and display mastery of grammar. Finally, to get a 6, Sam required that the essay display not only extensive content, going well beyond the basic requirements of the prompt, but also originality and a clear command of the language.

In examining the process by which Sam constructed his rating criteria, we can identify several distinct steps. Sam began with the key assumption, discussed under Research Question A that writers could be placed on a developmental trajectory. He then assumed that this trajectory described not only linguistic but also intellectual development. Postulating the existence of a developmental trajectory enabled him to not only estimate a writer’s level of proficiency but also her future potential, while the connection drawn between proficiency and intellectual development enabled him to compensate for situational factors by greatly broadening the range of specific indicators of particular levels of development. In subsequent steps, discussed under Research Question B, Sam inferred that the test takers were university applicants, that the purpose of the test was to measure “a student's readiness for university,” and that the role of the writing prompt was to elicit a language performance. These assumptions, combined with an evaluation of the essay topics that the writers he was assessing responded to, then helped him set performance expectations in general: for example, that potential university students should be able to cope with poor topics, or that linguistic competence was more important than task fulfilment. It also helped him to define a pass/fail threshold.

Turning to background factors, Sam's mutually reinforcing experiences as a learner and a teacher in the same educational system played the key role in both establishing a trajectory and relating proficiency to intellectual development. It is these experiences that, combined with social attitudes to language in his native country, led him to place a premium on correctness:

Grammar is very important ... even in my first language education grammar played a huge role in the assessment of my first language performance, and of everyone else's. Grammar was explicitly taught ... starting with grade 4, up until, I don't know, university. So that's educational background, and cultural background also.
This emphasis on correctness was initially prompted by Sam's successful language learning experience in a system (as measured by his performances on proficiency tests and by his experiences as a graduate student in North America) where formal aspects of language were heavily stressed. They were then reinforced by his teaching experience, which he said taught him that the mastery of formal aspects was a transferable skill which helped the acquisition of other areas of communicative competence in turn, such as the mastery of a range of genres, registers, and idiomatic expressions, as well as fluency (thereby becoming a reliable indicator of potential as well). Finally, grammatical competence was also a sure sign of logical thinking, and, by extension, of a high level of intellectual development.

A second important general sign of intellectual development, namely, the ability to generate ideas and organize them in a logical manner, was more evident from the frequency of Sam's comments concerning length, ideas, and text organization in his verbal protocols than from his statements in the interview. However, he made it clear in the interview that, in his opinion, students lacking the ability to gather and organize ideas would also have a hard time succeeding at university:

This person didn't have time and I am elaborating that they probably thought a lot. Because I have seen people do that [...] They're so focused on structuring their essay before they start writing it that they actually waste a lot of time doing that and then they can't finish [...] I think it comes from insecurity. Uhm, well, I expect an intelligent person to come up with those arguments pretty quickly [...] just the very fact that they have spent so much time on thinking up arguments, it's indicative of their, I don't know, intellectual abilities, or experience, which I would expect to be high for university applicants.

As for defining adequate topic development, Sam relied on his experience as a writer, as well as on "prescriptions" for TOEFL essays that he followed in coaching students, rather than on formal definitions:

There are some boundaries, that to me a good essay should, uh, fall into, and under a certain minimum it's not an essay any more; it's just some words on a piece of paper. ... I guess here my experience as a writer is the most important. Because as a writer, I have developed certain standards of how long it takes for a topic to be ... addressed. ... I have never been taught how long an essay should be. So, it's more of a feeling rather than an objective criterion.

Finally, less important, but worth considering for Sam, were neatness and handwriting. Although they did not directly reflect intelligence, they did, in his view, provide a measure of
writing experience, and since Sam usually equated experience with schooling and saw schooling in turn as the key to intellectual development, neatness and handwriting emerged as additional indices of readiness for university education, rooted in Sam's teaching experience:

Things like that, no lining at the left hand, it's not indented, ... show an inexperienced writer to me. ... Then "the lines are not filled": that's again very indicative of the writing habits of a person. If they don't know [they have] to fill the line and they finish the writing on that line ... it's very telling of, not ability, but the writing education that this person has received. After, you know, that many years of teaching ... you can distinguish between immature and mature handwriting, so that's why I would take that into consideration, if I see the handwriting of a person.

It is important to note that the same background factors which induced Sam to focus on mastery of grammar, quantity and quality of argumentation, and even handwriting as key criteria can explain his disdain for other elements of communicative competence that, in his opinion, either did not indicate intellectual abilities or even showed a writer's intellectual limitations. The most important of these criteria was fluency, frequently prized in North America, but frowned upon by Sam who felt it was easy to achieve in an immersion context and, in any case, did not guarantee success at the university level:

Fluency tells me that "this is a person who uses the language, as opposed to learns," but to me that's not a very relevant criteria, I mean relevant to the, the grade. To me that ranks somewhere around the bottom ... of all the criteria that I have used with them ... because my experience tells me that fluency is not, doesn't play a huge role in academic performance, because I have seen. It's like native speakers. They are fluent and their academic performance still can be terribly bad. So that's why to me fluency is no criterion of academic achievement.

According to Sam, then, the guiding principles for his judgements arose principally out of his experiences as a student of ESL, rather than EFL, reinforced by his teaching experience. He attributed some criteria, however, to his cultural background. For example, a key requirement for getting a score of 6 was creativity, and Sam felt that his appreciation of it was a form of opposition to the enforced conformity he had experienced in the political system of his native country. Another response to the political climate of Sam's youth emerged in his dislike of colloquial expressions which recalled for him the politically motivated destruction of academic standards in his native country.

Sam also drew a contrast between what he termed "Eastern European" and "North American" approaches to rhetoric and argumentation:
Our idea there of an essay was not this. It was more open, I would say. I didn't have that idea of a rigid structure, and, uh, a rigid way of developing an argument, rigid, rigid requirements, that you put your general ideas, and you know, topic first, general ideas or argument, then you develop them one by one, start with topic sentence, paragraphs start with topic sentence, things like that. We didn't approach it that way. It was more, the, I think the, the emphasis was more on cohesion, uh, cohesion, cohesive and coherent writing. Those were the two things that were the most important. Uh, after would come, I guess, style which is choice of vocabulary, and, uh, completion, I would say. ... Uh, clarity, of ideas. And, what else? I think, that's it. Uh, the general ancient Greek attitude to writing: it has to have a beginning, a middle and an end, but not exactly what that beginning middle and end should be.

Since he also held that a person with general writing skills would quickly master whatever style was required, and his experience with test-taking suggested that adherence to a particular style was not actually a requirement, he adopted a relaxed attitude towards the mastery of specific rhetorical conventions:

I think it's very normal for people coming from other cultures to have different approaches to writing structure, to rhetorical structure, and they shouldn't be penalized for that, and I think that once, for example, they enter a particular academic discourse community if they have these well developed ... general writing skills, they can be very quickly incorporated into that new academic writing discourse. Uh, I myself didn't have any problem when I was taking the TOEFL test. [...] And I had absolutely no preparation [...] I didn't know what those essays should be and I got the highest possible result. ... So that shows that even without knowledge of the particular requirements I was able to do what I was supposed to do.

Finally, based on his cultural background Sam drew a strong contrast between how East Europeans and North Americans constructed arguments, reflected by his comments on the perceptions of North American raters to an essay written by an East European student:

I remember [ESSAY #128] because I noticed that most people gave it much lower scores than I did, and I was trying to decide why. And I went over and over the essay again and again, and, uh, I don't know, to me it is still a good essay. I know it doesn't follow this rigid structure, it's very much unlike the Western writing and I guess I gave it that high score because that type of writing is very, uh, familiar and very, uh, close to I guess my own writing and my own experiences with the students. In terms of everything. In terms of style, in terms of structure. [...] Yeah it was very familiar. It came natural to me that the way that it was structured, the way that it went on ... in the beginning I didn't know it's, it's Ukrainian, but it, at the very beginning I liked it so I think that it's, it's very subconscious. [...] But no one gave it a 6, I don't think anyone gave it a 6. Uh, and there were like things like even 3 or 4. And I remember comments that it was presumptuous, and, uh, unwarranted, and almost things like that. ... I know why it's typical of East European writing. I don't know whether it was historically at that time. Our writing was developing in such, in such manner that you wouldn't expect data, you wouldn't expect hard data, uh, to warrant the arguments. What was more valued was the ability to come up with a generalization. So that's what, I think can see in this one, I can see in my own writing, and I can see it in the writing of the many Russian people that I have tutored. And when you confront them and say "You are just stating that. What is your proof?" they, in the beginning they cannot even get it that there should be proof. So for them, the very fact that you come up with a generalization, with a conclusion is much more important.
4.5 Summary

The developmental trajectory which Sam constructed drew on a wide range of experiences, although his learning and teaching experiences in his native country emerge from his interview as paramount to his establishment of scoring criteria, supported by references to his cultural background. These background influences were occasionally supplemented by prescriptions related to the TOEFL test, and by his intuitions as a writer. The absence of any formal assessment experience, as well as of explicit theoretical influences, is equally conspicuous, although it has clearly not prevented Sam from establishing consistent rating criteria.

Needless to say, any trajectory requires a beginning and an end, and in Sam's case the former is represented essentially by the absence of proficiency, while the latter is exemplified by his own performance on English composition tests which, although Sam generally dislikes norm-referencing, is one norm that he clearly uses:

I also compare people to my writing ... not consciously. I am not doing it on purpose but I do it. ... I can't say that I would compare any of the essays that we had, that I rated, to my own writing, because I think that my own writing is much, is superior to even the 6s that I gave. ... I am aware that my writing is good, and I am aware that I am better than most of the people who take the TOEFL test but still I have it, something like a standard in my head, that "that's the top." Probably, I think of myself as "that's the top that an ESL student can reach," O.K.? And there is a margin that I can still think of people approaching my level as the, the general one.

The final link in the chain of reasoning is formed by his perception, arising from personal experience of writing the TOEFL as well as from preparing students for it, that the TOEFL is a test for university applicants, or, in Sam's terms, of "a student's readiness for university."

How such expectations concerning test-takers and test use, combined with his ideas of proficiency and development, have helped him to set specific criteria is neatly encapsulated in his comments on how he established the pass/fail threshold for the corpus of compositions he was rating:

Teaching experience (TELLS ME) that people with particular kinds of problems in writing are at a stage where [...] their development would take a long period of time, and the other kinds of language problems I could say that students with that type of language problems develop fast. So I think certain problems, for example, uh, certain kinds of grammar problems [...] I can ignore. For example, if people have those morphological errors, (BUT) if the writing is good, if it's (OTHERWISE) either 4 or 5, but they make, for example, mistakes with idiomatic phrases, I don't consider that pretty serious. [...] and I think that problems with idiomatic phrases come from EFL learning and once they are in the environment, they will pick up the right phrases
pretty quickly. [...] But, if, for example, they make mistakes with adverbs, O.K., this is a pretty easy thing to me, because it's logical, it's easy to comprehend the difference between an adjective and an adverb [...] there is a problem with their intelligence. And if they are not able to do that, how will they excel in their academic writing in university? There will be a problem.
Chapter 5: Chris

Chris was a graduate student in his 30s, and a NS of English, raised and educated in North America. Besides having 6 years' university-level teaching experience by the time the study was conducted, his experiences in evaluating ESL writing included not only classroom and portfolio assessment, but also placement testing of students based on their writing abilities. He described himself as a "competent" rater. I chose him for the study partly because he was one of two participants in the Cumming et al. (1998) study with a science background, and partly because of the specific nature of his assessment experience. This experience involved extensive classroom assessment and placement testing, but not standardized tests, such as the TWE, administered on an international, rather than institutional, scale. Such experience, plus a limited familiarity with rating scales, placed Chris in an intermediate position on the dimension of assessment experience in the present study and helped me to select a full spectrum of raters of ESL compositions from relatively inexperienced to highly experienced.

5.1 Research question A: In what principled way did the raters of ESL compositions participating in the study relate the performance embodied in a written text to a writer's level of proficiency, and how did background factors influence this process?

In establishing the meaning of a writing performance, Chris didn't consider a writer's potential in awarding scores; his protocols contained no references to distinctions between scores for compositions and scores for their authors. Likewise, while he occasionally inferred that a writer ran out of time (e.g., #24, #62), he made no allowances for situational factors in his rating. A clear indication of his attitude may be found in his protocol concerning an essay (#105) which abruptly terminated after an introduction containing two perfectly formed sentences:

Next is essay #105. Right away I can tell it's going to be a 1. It's two typed sentences [...] what's written here is grammatically correct, as far as I can see, so either the person didn't understand the task, or they seem not to be interested in writing the [test]. So, anyway, a score of 1; there is nothing here.

In notable contrast with Sam, for example, who gave a similar essay (#126) a 2 based on his assessment of the writer's language control, even in the absence of most of the structural
elements of an essay, Chris considered only what he saw; he ignored what a writer may have been able to do under other, possibly more favourable conditions, and made no attempt to assess her future potential, either.

Chris' explanation of his attitudes to assessing writing performance focused partly on the concept of fairness, and partly on his classroom assessment practice, where he evaluated the product regardless of what he knew about its writer:

I guess it goes back to the theme of fairness, that I, it's not for me to second-guess what's on the page. [...] That's my belief, that if my task is to rate what's in front of me, that's what I do. And, also, in my classroom teaching I really make a concerted effort to do that, too. Not to consider, uh, "so and so, oh last time the essay was terrible," and come into marking a new piece of writing from that person with the same point of view. I really try to look at the product in front of me and judge it on its own merits. [...] I don't want to read something and (SAY) "Oh I know so and so can do better than this, so I'll give them a (HIGHER GRADE)." No. It's whatever is produced on that day.

However, striving to evaluate only what was in front of him did not prevent Chris from estimating proficiency from performance, a tendency that is clear from his interviews, if not from his concurrent verbal protocols. In discussing his rating criteria during his interview he routinely associated certain textual characteristics with certain levels of proficiency, as in the following comment on a writer who repeated an essay prompt in the opening sentence of her composition:

Again, from my teaching experience, I find that is a strategy often used by, when I say often, it really is often used, by less proficient writers, when they have nothing else to say [...] I don't, in my teaching, I don't completely discount it as a strategy, [BUT] I certainly ask them to avoid it as a strategy.

Such associations, relating a wide range of competencies and behaviours to a writer's level of development, signalled that Chris also related performance to proficiency through the establishment of a developmental trajectory from which specific rating criteria could be derived. That he never explicitly described his rating strategy in such terms, I would, in turn, ascribe to the nature of his assessment experience. In classroom assessment, which formed the bulk of his experience, he didn't have to estimate a learner's proficiency from every composition, since he could assume that it would "emerge" through on-going assessment involving not only tests but class participation, portfolio work, and so on. In addition, Chris
worked in an environment where the assignment of grades was secondary to giving feedback to writers; there was, thus, little need to ponder the meaning of an isolated performance. Although his experience with placement testing did involve the estimation of a writer's proficiency from a single performance, it was a low-stakes procedure in that "errors" in assessment could always be adjusted subsequently, by putting a misplaced learner in a different class:

You know, in placement testing it's never, uh, in stone. If that person happened to have had a bad day, they were not feeling well, then that will become apparent in the first day of class and they can be moved to a different class (FROM THE ONE THEY WERE PLACED IN BASED ON THEIR PERFORMANCE ON THE PLACEMENT TEST)

However, since the results of a test like the TOEFL are irreversible, even raters like Chris, who would normally either not assign grades in their assessment practice, or would do so under little pressure, would have to rely naturally, if not always consciously, on estimating proficiency from aspects of a single performance.

In spite of lacking practice in standardized assessment, however, Chris was able to develop consistent rating criteria, and, judging from his responses in his interview, he drew his associations of certain levels of proficiency with certain aspects of a written performance from two principal sources. First, his teaching practice enabled him to observe the performance of students at varying levels of proficiency, levels which he could estimate independently through in-class observation, portfolio assessment, and so on. Second, his assessments in placement tests involved matching students' writing skills with a curriculum for ESL composition instruction at the institute he was working for, and this curriculum also contributed to Chris' ideas of a developmental trajectory:

In rating, uh, in terms of looking at a set of essays, I have had quite a bit of experience in looking at essays and for the purpose of placing students in different classes, and we didn't really have a set-out (SCORING RUBRIC, BUT) we had some guidelines for what sorts of things should be evident in for each level.
5.2 Research question B: How did raters of ESL compositions interpret the role of writing tasks in performance-based assessment, and how did background factors influence this process?

Chris not only viewed essay prompts as specific tasks to be fulfilled, but frequently referred to them as "questions" which demanded an answer. To essays failing to "answer the question," i.e., address the topic and take a clear stance, he normally gave a 1, as the following summing up of the qualities of essay #140 demonstrates:

Uh, there seems to be a lot of — I can only assume they are typing mistakes, but who knows — uh, spelling mistakes in this. Uh, although the one sentence answers the question, none of the others do and there's virtually no development so this would receive a rating of 1.

Conversely, Chris gave essay #3 a 5, the highest score that he awarded, once he satisfied himself that the original question had been satisfactorily addressed:

Uh, just looking back over it … scanning the first paragraph again. [...] I just want to look back at the original topic. [...] O.K., uh, the reason I looked back at that, I was trying to remember if the topic had asked [...] in which case I was thinking that, perhaps, in the second body paragraph this person talked too generally. Uh, however … the person has addressed the topic quite well, quite effectively, I like the creativity in this. And, the brief conclusion, I think, is quite effective. There are a few grammar problems, nothing major. Mmm, I would give this essay a … 5.

In addition, as part of his evaluation of the compositions, Chris assessed task fulfilment (JR2) and the relevance of a writer's arguments to the assigned topic (JR3) 23 times each, such behaviour making up 3.4% of all the coded behaviours in his protocols.

In explaining his attitude to prompts in his interview, Chris referred to his experience with classroom assessment, which told him that topic avoidance was frequently used as a test-taking strategy by ineffective writers. While he stressed that outside the context of an assessment he encouraged students to write on topics that were relevant to their interests, he indicated that he was aware of the special circumstances surrounding a proficiency test, where topic avoidance, in his experience, usually reflected an attempt to recycle a memorized essay by a writer lacking in proficiency who hoped to improve her score. Such an attitude underscores the conclusion I arrived at in the previous section, that Chris was attempting to estimate proficiency from performance, and regarded topic avoidance itself as an indicator of a certain level of proficiency.
Related to Chris' demand that writers stick to the topic was his feeling that the four prompts to which writers in the corpus he was rating were responding to were unproblematic. Although he did read each of the four topics at the outset of his rating session, "to get an idea of what the tasks were," he did not analyze the specific challenges provided by the prompts at that time, and only referred to these challenges twice in the rest of his rating session. By contrast, he demanded task fulfilment and relevance. He confirmed the unproblematic nature of the prompts in the interview; although, like other participants, he deemed it necessary to check prompts for biases, in this case he didn't find any. This affords a notable contrast among the participants in the study: although all of them assessed the prompts for potential biases, the two NSs found the four essay topics used in the study to be neutral while the two NNSs considered some of them to be discriminatory since they addressed issues that were outside the personal experiences of socially disadvantaged writers.

A final reason for expecting task fulfilment and downplaying bias in writing tasks was Chris' feeling, expressed in the interview, that one measure of creativity was a writer's ability to stick to a topic on the one hand, and apply it to her situation, on the other. Consequently, knowing that the TOEFL was a test for university entrance, and feeling that potential university students, the test-taking population, should be creative, based on his own experiences at the university level, Chris was able to downplay the gravity of cultural bias in prompts in general:

I didn't feel having to write on [TOPIC A] was culturally biased, or insensitive. I think students read into a topic and they apply it to their situation automatically anyway.

Besides making the crucial assumption that the test takers were potential university students, however, Chris appeared detached from the writers of the compositions. While he commented on geographical names in compositions when they referred to places familiar to him (#21, #37), he not only never speculated about writer's ethnicity, language, or culture, but consistently overlooked explicit references to these in the compositions (#40, #128). Such detachment he attributed to his experience with placement testing, where he was instructed to ignore the ethno-linguistic background of writers in placing them in classes. This experience
contributed to a clear distinction in his mind between classroom evaluations and other forms of assessment in terms of the degree of familiarity between raters and writers:

If I have a personal relation, you know, students in my class, certainly their background is relevant, but if my task is to rate some essays, I see it as fairly irrelevant as to where the student has come from. Maybe interesting on the personal level ... "Oh, it's, I've been there!" but. Although I may not have commented on this in my think-alouds, I don't see it as relevant, as having any relevance whatsoever as to rating a piece of writing.

Such a distinction is a recurring theme in Chris' protocols, and reverses the attitude raters (consciously or unconsciously) assume towards assessing proficiency: in classroom tests, as discussed above, raters can avoid assessing proficiency and focus strictly on performance, while in standardized assessment they must estimate proficiency; conversely, raters should keep writers' ethnic and social backgrounds in mind in the classroom as an aid in their teaching through personalizing instruction, but must ignore them in standardized tests. It is significant that all four participants in the study followed both principles, even if only those with extensive experience with standardized assessment (Jo and Alex) fully appreciated the first of these contrasts.

5.3 Research question C: What reading strategies did raters of ESL compositions establish (to deal with both individual compositions and a corpus of compositions) and how did background factors influence this process?

Chris' macro-strategy for rating the corpus of papers was clearly formulated and consistently followed. First he read the essay prompts (although, as we have seen, he did not analyze them) "to get an idea of the tasks." Then followed a warm-up session where he evaluated some of the shortest essays in order to become comfortable with the rating task. Finally, he assessed each essay in turn, taking only one pass over the entire corpus, recalling, again, the "read-through" method of assessment (Milanovic, et al., 1995).

This procedure emerged from a combination of Chris' assessment experience with his interpretation of the instructions (discussed in Chapter 1) that he was presented with at the outset of the rating session. His experience of assessing placement essays in a university setting paralleled the current situation in that he had no prior knowledge of the proficiency of
the writers he had to assess. At the same time, one of the instructions to raters in the study which generated the protocols was that they should rely on their own judgements, rather than on the usual scoring rubric for TOEFL essays, in assessing the compositions. Together, Chris felt, these factors explained his practice of pulling out some of the shorter compositions at the outset of the rating session:

Because I didn’t have a rating scale, and was sort of constructing one as I went along, I thought “well, at least if I pull ones (HAVING) one or two sentences, they are likely to be in the same category so at least I can set my lower end, uh, anchor.” To relate it to my background again, in placement rating that’s something that we tended to do. Pull. Students were given half an hour, it was half an hour at that time to produce to something, and if all they have been able to produce in that time is one or two sentences, then just pull them right away and put them in a pile and there you have level 1. You know, of course, we read them through to, to verify that, but, uh, it again, anchors the whole set, the bottom anchor.

An additional reason for pulling out a few shorter papers was to make Chris feel comfortable — not with the rating task, which was familiar enough, but with providing think-aloud protocols, which was a relatively new experience for him. However, while Chris would rearrange papers even in the context of classroom evaluations, he made it clear in the interview that this was not a factor here:

I am entering my 9th or 10th year of teaching and I have taught a lot of writing classes and I would say I never mark, if I know the students, I would never mark them in the order in which they happen to be in the pile. I would always use some criteria, whether it’s to first mark some of the ones that I know might be stronger than others, if it’s towards the end of the session. But for this task, I didn’t, I don’t know, as I said I just pulled ones so I could feel comfortable with the think-aloud part, not the rating. And I didn’t know the students. Uh, as I said, that would reflect what I would do in a placement testing situation, you know, pull some of the shorter ones. In a classroom situation, no, because I know the writers, so it wouldn’t be reflective of that.

The instructions to raters also suggested to Chris that a single pass through the corpus would be adequate since the grades actually assigned were of limited concern to the study, and carried no practical consequences for the writers. So, while normally he would have read essays several times and sorted them into piles before assigning a grade, here he found no need to follow such a rigourous procedure:

(TAKING A SINGLE PASS THROUGH THE PAPERS), again, I think, was related to the project itself, that the instruction was to rate them given no criteria. And, uh, although it had crossed my mind to try to sort them and what not, uh, ... because I knew the rating was fairly arbitrary anyway, I didn’t see much of a point in doing that. [...] In placement testing, sure, you have to, because you have to get an idea of the whole group before you can start to compare and, oh definitely. And even in marking a class set, uh, I read them all through and read them again to
make sure that I am being consistent and fair in assigning grades. But for something like this, I just didn't bother.

Chris' micro-strategy of reading individual papers was likewise consistent: it began with a quick assessment of overall length, followed usually by a search for topic sentences. Then Chris read the essay and provided a running commentary, attending particularly to task fulfilment, organization, and logic. He edited errors at the same time. He gave no interim grades; instead, he concluded each protocol with a brief summation, often with a partial rereading of the text, and the assignment of a grade. That he closely adhered to this pattern is shown by a comparison of the first and penultimate essays, at opposing ends of the rating scale:

<table>
<thead>
<tr>
<th>Essay #140</th>
<th>Essay #21</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1</strong>: The first one in my pile is #140. I am just looking at the topic sentence.</td>
<td><strong>Stage 1</strong>: The next essay is #21. It is typed and two paragraphs long.</td>
</tr>
<tr>
<td><strong>Stage 2</strong>: [QUOTE] There are several typing or spelling mistakes, it's hard to tell with a typewritten sheet. [QUOTE] Here the student clearly answered the question, in that sentence anyway. [QUOTE] Non-idiomatic English there. [QUOTE] I guess, there is a typing or spelling mistake. [QUOTE] And then the essay stops.</td>
<td><strong>Stage 2</strong>: Nice, clear beginning. [QUOTE] Uh, &quot;ridiculous&quot; is spelt wrong. [QUOTE] OK, not clear who &quot;they&quot; are. [QUOTE] Ugh! God forbid! [QUOTE] And there is a comma splice. [QUOTE] OK, although there are some structural problems and some pretty major grammar problems such as the comma splice, uh, the first paragraph I find it kind of funny and the tone is fun, a little bit creative. Next paragraph. I am thinking that, perhaps, the student was living in [...] because the next paragraph begins with [QUOTE] Looks like a typing mistake, it couldn't possibly be a spelling mistake. [QUOTE] Good vocabulary, and entertaining story, so far. [QUOTE]</td>
</tr>
<tr>
<td><strong>Stage 3</strong>: Uh, there seems to be a lot of. I can only assume they are typing mistakes, but who knows, uh, spelling mistakes in this. Uh, although the one sentence answers the question, none of the others do and there's virtually no development, so this would receive a rating of 1.</td>
<td><strong>Stage 3</strong>: OK, although it's, you know, the first paragraph is entertaining, and also it's very creative, uh, sort of seems a bit off, a long way around to getting to the actual story itself, but I have to admit I like it. I like the creativity and the tone is, as I said, fun. The second paragraph, describing the ice storm, good use of vocabulary, pretty entertaining writing, certainly not boring. A few struc-, grammar problems, the ending is excellent. I would rate this one as 4.</td>
</tr>
</tbody>
</table>

In his interview, Chris related his practice of reserving the assignment of a specific score till the end of his rating to his general preference for feedback over grades in his teaching practice, most of which involved non-credit courses:
I think if I need to relate that to my background, uh, most of the writing classes I have taught, I have not been required to submit any grades for students. ... So, that may, it is just occurring to me now as I am reflecting on it, that may be a very big part of why I am not grade-oriented as I read through a piece of writing. I am interested in the writing itself, not attaching a number to it. And that relates directly to the, my past experience because that is the context that most of my writing classes have been.

Chris explained another aspect of his strategy, the assessment of length at the outset, which was not normally a part of his procedure in classroom assessment or placement testing, with reference to the instructions he received before rating his corpus of compositions:

I think at the beginning of each think-aloud I tended to look at the overall appearance and I think that is just because these were the instructions for the think-aloud protocol. Uh, I don’t think that this is something that necessarily comes naturally to me, that I would naturally do. I would just start reading.

Finally, besides making no references to norms, either external or arising from the corpus, Chris was the only rater in this study who did not assign a single 6 in his rating session. And, although he did not say so in his interview, the nature of his assessment experience may also explain his avowed dislike of norm referencing: placement tests aim to match a writer's existing skills with specific courses and in such a situation a writer's relative standing in a group is of little relevance; it is, instead, her mastery of specific skills that is relevant.

5.4 Research question D: By what process did raters of ESL compositions construct specific rating criteria, and how did background factors influence this process?

Proportional frequencies of Chris’ coded behaviours presented in Figures 7 and 8 demonstrate that he focused on rhetorical-ideational qualities to a much greater extent than on language throughout his ratings, particularly when making judgements (see Figure 8). In common with the other participants, in the initial stage of his readings of individual compositions he commented mostly on length and appearance. Then, during the running commentary stage, his interpretive comments focussed more on language (see Figure 9) while his judgemental comments focussed on rhetoric and ideas (see Figure 10). In the final, decision-making stage, where he actually assigned scores, his interpretive and judgemental comments both focused heavily on rhetoric and ideas (see Figures 11 and 12).
Figure 7: Proportional frequencies of interpretive strategies during Chris' rating session.

Figure 8: Proportional frequencies of judgemental strategies during Chris' rating session.
Figure 9: Proportional frequencies of interpretive strategies during phase II of Chris' rating session.

Figure 10: Proportional frequencies of judgemental strategies during phase II of Chris' rating session.
Figure 11: Proportional frequencies of interpretive strategies during phase III of Chris' rating session.

Figure 12: Proportional frequencies of judgemental strategies during phase III of Chris' rating session.
Further, within the realm of ideas and rhetoric, Chris placed notable emphasis on "reasoning, logic, or topic development," which made up half of the judgemental comments he uttered on rhetoric and ideas, and outnumbered all the judgemental comments on language put together. In addition, there was a relationship between Chris' assessment of "reasoning, logic, and topic development," on the one hand, and the score he assigned to a composition, on the other. Essays Chris rated 1-3 received overwhelmingly negative comments on “reasoning, logic, and topic development,” while those he rated 4 or 5 received mostly positive comments.

Additional rating criteria extracted from Chris' comments in the protocols include the presence of the key structural elements of an essay — an introduction containing the thesis statement, body paragraphs introduced by topic sentences, and a conclusion — as a minimum requirement for lower level writers; "adequate" topic development (including both a certain quantity of information and a degree of coherence) combined with a clear stance on an issue; "academic" tone, which is defined largely negatively as the absence of dramatics such as the use of questions to introduce an essay (#62), and the avoidance of sloppy expressions such as "etc" (#88); and creativity, both in ideas and, to a lesser extent, in language, all of which Chris regarded as prerequisites for getting a score of 4 or higher.

Conspicuously absent from the list, of course, is language control. This is interesting since Chris actually noted errors frequently, as shown by the fact that 23.7% of all his interpretive comments concerned language (Figure 7). However, in rendering judgements, Chris usually assessed the mastery of formal aspects only when he needed to decide on borderline cases. In the interview, he summed up his attitude thus:

I am noticing a lot of grammar mistakes; I can't help but notice them, uh, as I am doing any kind of marking. But as I said, it depends; for these ones I wasn't marking from any kind of a rubric, uh but in my own rubric, in the rubrics I tend to use when I am rating my own students' essays [...] I do have a category for grammar but it's an overall thing and it certainly isn't weighted very heavily.

In establishing such rating criteria, Chris relied heavily on his teaching experience to construct a developmental trajectory that established what students at certain levels of
proficiency could do; for example, that creativity or mastery of a range of sentence types were indicative of high-level writers, and redundancy of relative beginners. Just as important, teaching experience told him also what students at a certain level would be unable to do, and how they could strategically compensate for their shortcomings:

I find that often in [THE ESSAY OF] someone who is a very proficient writer [A THESIS STATEMENT NEED] not be stated directly but generally I find with second language writers that, uh, they are not at that level to be able to do that. So there needs to be an explicit statement, either a Topic Sentence, well ideally a Thesis Statement or a Topic Sentence, so that their essay keeps on track and has a clear purpose. [...] I have, at a very advanced level I have seen people who are able to carry it off, but these, particularly these, this batch of TOEFL essays, were generally not very good overall.

Indeed, his teaching experience even enabled Chris to identify the competencies that students needed to master in order to move to a higher level of proficiency:

In general I find that [TOPIC DEVELOPMENT] is the biggest problem. Usually by the advanced level, which is what I was teaching at that time, they know the basic format of an essay and we are working on more stylistic things and development strategies.

Other criteria that Chris directly attributed to his concurrent teaching experience included a clear stance on the assigned topic in the thesis statement, relevance, and directness. Above all, however, it was his teaching experience, where he seldom saw writers transferring their grammatical knowledge to written communication, that made Chris question the value of focused grammar instruction; consequently, he was able to subscribe to a basic principle of Communicative Language Teaching (CLT), the received wisdom in the department he was teaching in, that it was in the effort to communicate that grammatical structures emerged:

I think attention to grammar is important, and, and I think that it is something that can be taught and learnt, [but] I believe it's more useful to learn it in the context of whatever you are working on, in the context of writing, and editing your own writing. But I think it does eventually come, because I think if you, as a grammar teacher if I teach a grammar point and I teach the same students writing, I don't see the transference from grammar class to writing class. However, I do think if they do enough practice in writing, that they start to form the rules in their head, and for me that is the ideal way to learn grammar. ... in general, in terms of verb tense and ... vocabulary and structures, I think that they will, eventually [EMERGE IN THE EFFORT OF COMMUNICATING MEANING].

Note, however, that even though Chris downplayed the value of grammar instruction, and relied more on evaluating reasoning, logic and topic development, he acknowledged that mastery of grammar would actually be a good indicator of proficiency in his developmental scheme. This is suggested by the relatively high number of positive comments on language
use in papers Chris rated 3 or higher, as well as by comments such as the following, which he made during the course of the interview:

> If I see that a writer is able to use a variety of sentence structures, not just compound but complex or... you know, complex-compound, I think that shows their development in terms of writing proficiency.

> Needless to say, teaching experience itself does not arise in isolation; rather, it takes into account institutional pressures manifested in peer attitudes and curricula, as well as impressions gained from the theoretical literature. That Chris recognized such influences was evident from the following summary of the origin and essence of his teaching beliefs:

> Where most of my work experience has been it was the belief, hmm, the, the culture ... that we should teach writing as a process, we should focus on ideas, uh, that grammar ev-, will emerge, and that in writing class and in grammar class, you know you try to pay attention to certain forms and when the student is at a level, able to accept and understand that form, will start to incorporate it in his or her writing. So I would say that it was the prevailing culture, and I believed it myself, and I think have read enough in the literature to support it. ... And I just think in my teaching experience, as I said, I don't see the transference from grammar class into, into what they are doing. I can do a lesson on, I don't know on what, adjective clauses. OK? And they do a piece of writing, and adjective clauses are hopeless and, and then I can even say, "Remember yesterday we did this?, and they go "Ah!" and it, it just doesn't...Hmm, I don't think things can be taught in isolation like that, I really like an integrated approach to language teaching... and I like grammar errors to emerge from the writing and then we will try to address them in the context of their writing, and they may not get it the next, or the next or the next time, but eventually, I think, I believe, that people form their own internal rules. ... Yeah, and I never thought of it but it was the prevailing approach where I was working, a process approach to writing.

Such an attitude to grammar was not only incorporated into the curriculum, but also applied to assessing placement essays, another of Chris' responsibilities at his institution, deepening the process of enculturation:

> As long as the content was at a fairly sophisticated level, uh, the grammar and mechanical problems that may have been present were not factored into where to place them in terms of class level. Because if someone had fairly sophisticated thoughts, the thinking was amongst us that that person would more appropriately fit into a more advanced level class.

All in all, Chris' comments concerning the origins of his scoring criteria support Pula and Huot's (1993) thesis that enculturation in discourse communities can explain the behaviour of raters. Most of Chris' scoring criteria came from his teaching experience, which involved a general emphasis on communication and a process approach to teaching writing, both incorporated into the curriculum, as well as the use of certain rating scales. On the other hand, unlike raters in Pula and Huot's study, Chris relied much more on his teaching than on his
assessment experience for scoring criteria. In addition, rather than accepting CLT principles blindly, he used his practical experience to identify theoretical principles that actually helped him to explain his students behaviour. Consequently, while there is support for the enculturation process postulated by Pula and Huot (1993) through the agency of discourse communities, the case of Chris' rating behaviour shows that the process is clearly not a simple one.

5.5 Summary

Chris' scoring criteria relied on a developmental trajectory from which he could infer a writer's level of proficiency based on key indicators. The key assumption underlying his trajectory, arising out of his teaching experience, was that learners acquired grammatical structures through the effort to communicate, an assumption he operationalized in his teaching practice which focused on logic, topic development and essay structure. It was the latter that he particularly looked for in the essays he rated; although he frequently noted linguistic errors, judgements on language use played a limited role, except at the advanced levels of proficiency where creativity and sophistication became criteria for getting scores of 4 or higher.

Although he didn't set pass/fail thresholds, Chris did rely on norms to supplement his developmental trajectory. His practice of pulling out "good candidates for a score of 1" at the outset of his rating session suggests that he began his rating session by anchoring his scale at the bottom. That he had already established other scale points prior to the rating session is shown by his statement in the questionnaire that he used the proficiency of an educated native speaker as the norm for a top grade; in his interview he defined this norm thus:

I think native-like competence as defined in the, these rubrics I am referring to, is defined as that of a, I don't think, I think for politically correct reasons this is never stated, but uh, a fairly educated person for whom English is the L1, would be expected to write. [...] the rubrics that I am familiar with would refer to such things as "appropriate and varied use of vocabulary," "almost non-existent grammar errors," uh, "facility of expression." These are descriptors that I am pulling from other rubrics.

Such a definition of the top level of his developmental trajectory may explain why Chris did not assign a single 6 during his rating session, even to essays to which some of his fellow
raters did assign the highest score. Another way in which Chris anchored his developmental trajectory was by using the students he was concurrently teaching as a reference group:

After I had finished rating (THE ESSAYS), I think I tended to be, uh, a bit harsh, on, in some of them. And, again, I think that was because at the time I was teaching very advanced students, and so I probably, although maybe not consciously, uh, was thinking, you know because I was used to reading fairly, you know my students were pretty good, so probably in the back of my mind somewhere I was comparing them ... perhaps, I don't know.

Chris' feeling that he was harsh is borne out by the fact that out of the 70 cases where he and another participant in the present study rated the same paper, Chris awarded a lower score in 48 cases, an identical score in 17 cases, and a higher score in only 5 cases.

Finally, Chris also made assumptions regarding test use and test takers, even if he avoided envisioning the personal situations of individual writers. To wit, he assumed that the writers of the TOEFL essay were university applicants, and this contributed to the setting of certain expectations, such as the need to produce an academic essay, or the need to show creativity. At the same time, he expanded his notion of test use to include the fact that his ratings in the present study carried no practical consequences, since they were undertaken for research purposes. Consequently, he modified his rating strategies, reading each paper only once and not setting a pass/fail threshold.
Chapter 6: Alex

Alex was an East Asian graduate student in his 40s at the time of the study, and the only doctoral student among the 4 participants. I chose him for the present study because of the breadth and depth of his experience. In addition to classroom evaluation and placement testing Alex had worked as a composition rater for a nation-wide English examination authority in his native country and had also been involved in rater training. His teaching experience had been likewise varied, including ESL at the secondary and tertiary levels of education, as well as English for Special Purposes (ESP). He was the only participant in the first phase of the Cumming et al. (1998) study who described himself as an expert rater, and his rating strategies differed markedly from all but one of the other participants in that study. Being at the upper end of the continuum of professional assessment experience I thought he would be of considerable interest to a study of rater behaviour.

6.1 Research question A: In what principled way did the raters of ESL compositions participating in the study relate the performance embodied in a written text to a writer's level of proficiency, and how did background factors influence this process?

In his questionnaire Alex listed "knowledge and understanding of the relationship between proficiency and performance" as a key factor influencing his assessment of compositions. While he did not elaborate his views there, and gave few indications of his attitude in the verbal protocols accompanying his ratings, he was very clear in his interview that he associated a certain level of proficiency with a given performance. Among background influences, he cited his teaching experience as one of the major factors:

I do make use of my background knowledge, in terms of having some kind of a matching between the language in a writing and the language, the estimate of proficiency level, for example. I think many teachers, uh, I mean not necessarily raters, would have that kind of a knowledge or assumption. ... I have taught at tertiary and secondary level, and also junior secondary level, so I think I have experience with many different levels of learners, especially ESL learners and when you see a piece of writing you do have some estimate as to what level, you know, this writer could be, or should be. It's not a sure thing but you do estimate that knowledge [FROM] non-test situation, classroom situation, as well as tests themselves, because we do conduct tests in schools, especially in secondary schools, because of preparing students for these public exams, we do give them similar tasks to do. Uh, I have, actually I taught graduation level for many years, so, and I rated the same public exam for
many years. So, that kind of an outside-testing-context knowledge could help me to associate a certain performance with a certain level of proficiency.

Apart from his teaching experience, his familiarity with the Teachability Hypothesis, which holds that syntactic and morphological structures in second languages are acquired in a certain order (Pienemann, 1986; Pienemann, Johnston & Brindley, 1988), also led him to associate a certain level of proficiency with a certain type of performance, even though he was aware of the criticism directed at this hypothesis and didn't derive any specific rating criteria from it:

The acquisition of certain syntactic or morphological structures is stage-wise [...] Now that line of research, I think, although it's been challenged by more recent studies, it's still very much in the back of my mind in a certain way and that actually has formed a base, a theoretical base for the assumption that a certain performance is associated with a proficiency level. So to say that all this knowledge comes from my teaching is probably overly stated. I mean I do think that, uh, as a rater the use of, uh, various types of information and knowledge in rating may have sometimes come from the theory in the literature. And, although this kind of research has been challenged, I don't think, I think that there is some kind of gradation there in the acquisition of certain grammatical structures and I do believe that.

Conversely, while Alex explicitly envisioned a writer behind a composition in 19 cases during his concurrent verbal protocols, only once did he attempt to infer a writer's potential, feeling that her performance was not a "true" indication of her proficiency (#44). In the rest of his protocols, there was no indication that he was looking to infer a writer's proficiency from anything other than her performance, even in borderline cases.

In like manner, Alex occasionally identified situational factors affecting a particular performance; for example, he thought that the writer of essay #22 was under time pressure and that the writer of essay #13 wasn't, and he admonished the writer of essay #93 for writing in a hurry. However, Alex made no allowances for such factors and in this sense he did, indeed, rate compositions at face value, as is evident from, for example, the following summation concerning a clearly unfinished essay (#104), which he thought showed potential:

"There are some minor errors, but what little is said is basically clear. [...] Uh, I'll put a 2-plus for the time being. It's a little too short. So, kind of, this student, he may be able to write, I mean in terms of proficiency. But, obviously, there isn't enough content to judge. So, it could be a 2. Let's put down a 2-plus."

In explaining this tendency in his interview Alex stressed that the principal advantage of performance-based assessment was that "you cannot go beyond what the performance
suggests." Coupled with his assertion, cited above, that in performance-based assessment one always tries to estimate proficiency from an isolated performance, I can assume that by "rating a performance at face value" Alex meant trying to infer proficiency without taking into account situational factors or a writer's potential into account by speculating what she might have been able to do under ideal conditions.

6.2 Research question B: How did raters of ESL compositions interpret the role of writing prompts in performance-based assessment, and how did background factors influence this process?

In the questionnaire, Alex listed task completion as the key to success in composition examinations. This was clear from the attitudes he expressed in his protocols: while he was willing to overlook a degree of irrelevance if the writer's language was otherwise proficient (as in the case of essay #4), he consistently assessed task fulfilment. Further, he frequently equated task fulfilment with "answering the question," and in referring to essay prompts reduced them to simple questions. Last, but certainly not least, task completion played a major role in establishing an interim, 3-point scale which he applied to the compositions during the early stages of his protocols (see 6.3).

In explaining his emphasis on task fulfilment, Alex referred in his interview to a specific assumption concerning the candidate population for the TOEFL, one that he later recognized as somewhat inaccurate:

When I rated I had the assumption that these were students more or less ready to go to university. Now this is the case in (MY COUNTRY, BUT) later on I learnt from other colleagues that in some countries this is not the case. I mean, in some countries you don't actually have a prior screening of some kind. It's more like, "if you are ready to pay, you pay that fee, and you take the exam.

Armed with this assumption, Alex could construct specific expectations, based on his university-level teaching experience in his native country, where he followed a task-based syllabus that expected students to master a battery of specific skills involved in the ability to successfully complete a task:

I guess, in the later years of my teaching, uh, I shifted gradually from a functional-communicative syllabus to a task-based syllabus. OK, so, I do have this concept that task completion is an important criterion in the writing assignment. Task sometimes requires not just use of
language but also, uh, cognitive reasoning, real world knowledge use, that sort of things. [...] it could be a personal [SHIFT], because I stopped teaching secondary school which by and large used a functional-communicative approach [...] whereas in university, because I was teaching technical writing, it was more a task-based syllabus. [...] So, from that respect you no longer just look at language, language control and proficiency.

Alex’s attitude to task fulfilment also relied on his observations of students’ actual performances on the TOEFL, helping to further refine his specific requirements:

I also recognize that this writing is sometimes like a game. He plays it, I play it, I don’t have to believe that he really believes what he writes. I don’t have to. Some people will find an easy way out. Write something that they don’t have to think a lot. Look at the content, at the argument of this sort of writing, and you will know a lot of them, although a lot of them are logical, or well presented, or well supported, it’s something we would call a cliche, after all. It’s something that we would. Because the prompt in a way invites, especially if the writer doesn’t have the personal experience of that context, he could really write some pretty formulaic content.

It is true that Alex’s expectations regarding task-fulfilment were tempered by his experience of public examinations, particularly when standardized across a wide range of cultures, as well as by his specific knowledge about the TOEFL:

In this project, because I know this is TOEFL, and because I know that the task requirement is not very specific, it’s more like “you have a task because you want to give them the, some, some context to write something.” So, those aspects are not, I felt, at one point, not very important. So, I would still try to give some of these candidates a 3 or a 2, depending on their display of language. I probably may have given one a 4, knowing that he didn’t complete the tasks but still displayed a certain level of proficiency. [...] Now I understand that the contemporary writing researchers do seek to avoid asking display questions, but I don’t see how this is possible, being a public exam like TOEFL, frankly, I don’t see how it is possible. It is much more possible if we can contextualize it further down, the prompt, to, you know, individual level. But then the question of comparison comes, you know, like reliability. So, it is not a question that, I think, can be easily tackled in a public exam as large as TOEFL.

However, Alex retained task fulfilment as a key criterion in spite of his awareness of potential cultural bias in TOEFL writing prompts, a bias he downplayed by referring to the specific use that TOEFL scores in his experience were put to, a factor he had identified in the questionnaire as exerting a major influence on his assessment:

I understand, as I said earlier, that some of the prompts are not really applicable to different cultures, and different communities, but there is no way of avoiding this as TOEFL being an initial exam for North American universities. It’s not like TOEFL is an exam that anybody learning English in the world has to take, you know. So, in a way I would go along with that sort of a prompt, although I said it’s not ideal, but it’s fair enough because “you guys want to come over to North America to study, so my requirement is you take TOEFL.” [...] So I mean, no, I don’t think I would benefit a writer who I assume, which I suspect coming from a different culture, because there is no there is no reason for compensating that writer because first, it’s only a suspicion of mine that he is being disadvantaged because of the prompt. Second, it’s more like a voluntary exam, not like a public exam which everybody has to take, you know.
It is not surprising, then, that while Alex was not averse to speculating about writers' mother tongue (#22, #48), academic orientation (#74, #91), and even age (#5), his comments on writers' backgrounds, came in the form of asides, uttered during his initial pass through the compositions. Indeed, once (while reading #21) he even cautioned himself against speculating about writers' backgrounds, presumably for fear of letting this influence his judgement. Apart from his assumptions concerning test use and test takers, he also referred to the dangers of attachment in standardized testing in general, a danger that had long been recognized in the public examination system of his native country:

I guess, attachment in this situation is [...] an equity issue. It's not fair to certain writers. I mean I try to be detached, not because I think we should. I mean as a reader, when you read something you do want to seek contact with the writer, sometimes even in terms of personal contact. But, in assessment we try not to do that so that we won't be biased against certain types of candidates. The good thing about performance assessment is that you cannot go beyond what the data or what the performance suggests. [...] The origin of exams, especially in (MY COUNTRY) was to decontextualize candidates, so that their talents could be assessed in terms of the talent, you know, their writing ability, their eloquence. ... Not where they come from, whether they come from a poor village, or they are a farmer's daughter, uh, son. I mean, that's why you have the exam, right? This is testing. I think it's a very revealing remark, you know, like "we assess because we want to decontextualize other factors." So, what's the point of testing [otherwise]? So, uh, I do hold a more detached view.

6.3 Research question C: What reading strategies did raters of ESL compositions establish (to deal with both individual compositions and a corpus of compositions) and how did background factors influence this process?

Alex began his rating session by briefly familiarizing himself with the writing prompts, and by scanning a few essays for length and appearance. This was followed by an initial reading of all the essays for the purpose of assigning tentative grades (with + and - signs used in cases where Alex was unsure about even a tentative grade). During this stage, Alex frequently paused to set criteria for the various levels of his emerging rating scale, which involved repeated rereadings of some essays until the criteria were satisfactory. Finally, once Alex finalized his informal rating scale, he went through all the essays, now sorted into six levels, in descending order, and assigned final grades to each one. His approach corresponded roughly to a "principled two-scan read" (Milanovic, et al., 1995). However, while initially Alex read a paper through at least once and often repeatedly, in the final stage he terminated
his reading of a composition as soon as he felt confident enough to confirm or revise his original rating. An example of this procedure, followed consistently, is furnished by his protocols concerning essay #116:

(Initial reading) OK, I'm gonna begin with 111 now. (QUOTE) What's that word? Oh, (QUOTE) What's that, there's the missing "s." (QUOTE) Oh god, what's this word? (QUOTE) OK, uh, (QUOTE), I've just turned the page. (QUOTE). What's this? (QUOTE) I don't know; what's this word? (QUOTE) They can be, uh, that's strange! Only, it doesn't it doesn't look like, oh, OK, what you mean, they can be willing, you mean they can be and willing be done?... (QUOTE) OK so, this person is proposing (QUOTE). Uh, let me see. But this second paragraph, what's the central idea here? (QUOTE) OK this is assuming that the student will be able to [...] Uh, (QUOTE), oh. OK. (QUOTE). Well, OK, (QUOTE) It's a lot of assumptions here that haven't been defended. Now, this knowledge of course cannot be obtained. I am reading now, OK, (QUOTE) OK, well, [...] OK, I understand. OK, Oh, I missed that word before. (QUOTE) well, how short is short? How long is long? That is undefined. (QUOTE) That should be "latter," this is (QUOTE) Well, uh, this conclusion is sounds to me absurd although I can see that this student has quite good control of the language. Now, tentatively speaking, I'm supposed to ... I'm supposed to assign a number, let's see. Uh, what number should I assign now? OK, let me uh, come back to my sheet here, so, I'm supposed to assign uh, a number between 1 to 6 to each "based on your assessment of the quality of each." Quality, OK. What's quality means to me? Well, mmm. I think, yes, as I said he has full control, although the the argument itself is rather superficial and there are a lot of undefined assumptions and the argument uh I think uh, from the perspective of proficiency, I would, uh, let's say I would assign a 5 maybe. I would assign a 5 to this script, but I may change it later, let me put a 5 here next to 111. There are a few silly mistakes, but they are not, they are not really serious errors. OK, I think it is a 5 or a 4. I just put down a 5.

(Second reading) And then I want to go back to 111 and put a plus after 5. So, that's a 5-plus. By that this sign is to help me to determine later whether I would upgrade these scripts depending on the other scripts.

(Final reading) OK, 111. (QUOTE) There's a missing s there. (QUOTE) Uh, OK, well, he is good. Good rhetoric, good. (QUOTE) Uh, yeah, I hate this sentence. It's messy. It's very messy. So, (QUOTE) uh, is there any reason for me not to give it a 6? Short, uh, (QUOTE), OK, there are some minor errors, but they are not that important. Uh, (QUOTE) there there are some statements, statements in the essay that are not supposed to--. That is why it is a 5.

While Alex did not render interim judgements during the first reading, the existence of multiple readings meant that the grades he assigned at the end of the first reading were invariably interim, except for papers rated 6, which were never reread. The function of the second reading, in stage 3, then, was to test the hypotheses represented by the interim grades Alex assigned to a paper at the end of the first reading, although in practice he only modified grades in the case of serious discrepancies between an interim grade and a revised assessment.

Multiple readings and interim judgements thus represented two key elements in Alex's rating strategy, ensuring reliability: they were necessitated, in his experience, by the impossibility of attending to every textual feature in a single reading, and the impossibility of rendering a final rating until the entire corpus had been read through at least once:
That, that kind of thing happened a lot in the previous university where we had two examiners rating the same scripts. And, you know what we had to do? We had to, if it's 200 scripts that we each of us marked, we had to sit down and display our own assigned grades on a sheet and compare. If it's two mark difference, we go back to them and standardize again and read. And a lot of the time you do find that your own assessment is "Oh no, this is not really the number I want" So you can always go back to them and reassign them. You can't just think for sure that "if it's a 6, if it's a 6 it's definitely a 6." There is no way, I mean, the human mind works very mysteriously. Sometimes you are caught up more by one aspect of the writing, and you are obsessed with it and it's sort of upset you, your judgement, you know, and the next time you come back to it you decide to change your mind. That happens quite frequently. You know, you don't just give one shot rating, I mean.

In addition, since Alex believed that scores in a large corpus should show a normal distribution, he felt he needed to review his assessments at least once in order to make sure that the distribution of the scores was not skewed. Finally, unlike some raters in large-scale examinations, Alex did not use anchor papers to ensure consistency, and this also necessitated multiple readings.

Recognition of the need for multiple readings came, in Alex's opinion, from his assessment experience, which included 9 years' service for a public examination authority in his native country. Although in explaining his read-first-judge-later procedure he cited cultural values, such as respect for others' opinions, he felt that his experiences as a rater were paramount:

In [L1] we have a saying: "Idea comes first." OK? I want to know what this guy is saying, I mean to give it due respect. I mean my role as a rater doesn't come in right away. I mean I have to know, I have to play the role of a careful reader first. [...] This is how I approach a text in [L1]? I guess so, I really don't know. [...] I mean it's probably not a cultural thing, it's probably more a rater thing to me.

The assignment of interim scores likewise arose out of personal experience, although Alex felt that it was also congruent with the logical positivist approach to advancing scientific knowledge that he had once favoured; these two factors in combination allowed Alex to develop, and articulate, a consistent approach to rating:

I do give [a grade] towards the end [of the first reading], but that is more like, more like my first shot at it. And, you have to place it somewhere on a scale, so that you know whether you have done the right thing. I mean you sort of, you have to objectify your thinking, your assessment, by assigning an initial number and then you look at this initial number, because now it's objectified, outside you. [This is due to], oh, my background is in science, ... yeah, K.R. Popper, you know. Yeah in my early days I was a fan of the logical positivists. Not any more. But I guess the modern way is to say I objectified my thinking, my rating, by putting down a number and then come back and reassessed that number. It's not like, when I come back I reassess the writing. It's more like, you could say, metaphorically, that I come back and and look at my assignment, you know, this number, whether this number is valid or relevant.
The other outstanding aspects of Alex's procedure were his extensive reliance on norms in assessing the compositions, and his sorting of compositions into piles, a process which continued until the piles were internally consistent and clearly distinguishable from one another. Both of these practices arose naturally from Alex's experience with large-scale public examinations. Further, Alex used not only external norms in assessing the compositions (such as the proficiency levels of students he was concurrently teaching), but he also took the qualities of the corpus he was rating into account. He ascribed such a preference for contextualized rating to his general dissatisfaction with using "universal" marking schemes in his assessments:

Well, in this project I never used a descriptor, but in all the exams that I marked, that I rated, including institutional exams, in department, in school, we all use a marking scheme and in fact in later years, at the university level, we decided that a universal marking scheme didn't work any more. It was more like you use a [different] marking scheme every time for a specific prompt or task. I mean, devise a new marking scheme every time and you don't use the same one. We have this problem, because when we deal with different document types you realize that you do need different..., uh, it's better, I mean it's not like that [the other] won't work, but it's better, it's more reliable if you devise a specific scheme for that particular task.

What is of great importance here is Alex's feeling that however detailed they may be, scale descriptors are never adequate by themselves, meaning that they would always have to be refined taking the specific qualities of a corpus of compositions into account:

I think a descriptor is not helping much, in my view [BASED ON] past experience with descriptors for holistic rating. Let me put it this way. The problem with a descriptor is that, is the assignment of proportions of different aspects of a piece of writing. I mean they always cross, interact themselves. You know, you have to look at the interactional effects between and among the various aspects that eventually you come down with a simple number. So that while the descriptors are there as a guideline, more like, and they have no substantial help in terms of deciding whether [a paper] is a 4, because a 4 is sometimes short in grammar but strong in ideas and then you have organization and so forth. So what is a 4? A 4 can mean a host of longs and shorts of many aspects. ... So to me it's more like trying to do the impossible. It would be much easier to say "OK I am going to go by language, just language itself, and then line them up and maybe then I could write a good descriptor."

In addition, Alex felt that he could assume that the scores in a corpus of 60 compositions were normally distributed, and this assumption did help him to decide on borderline cases, independently of a composition's textual qualities, as in the case of essay #135 which was scored a 4, rather than a 5, because Alex felt that he "could afford to give more 4s than 5s" when using a 6-point scale.
6.4 Research question D: By what process did raters of ESL compositions construct specific rating criteria, and how did background factors influence this process?

In rating essays, Alex first created a rough, 3-point scale, giving interim scores of 1 or a 2 to "below average" compositions, 3 or 4 to "average" ones, and 5 or 6 to "above average" ones. He only applied a 6-point scale to the compositions in the second phase of his ratings. In establishing what constituted "average" Alex compared the compositions with what he thought students in the university writing courses in his native country could produce. He subsequently refined his scale in order to separate essays he rated 3 from essays he rated 4, which, given Alex's assumption of a normal distribution of scores, were expected to provide the largest number of essays in the corpus. If Alex thought he had seen the level of writing in a particular essay previously among university students, he awarded the essay a 4; if not, he awarded it a 3. Such holistic comparisons of essays to the levels of writing exhibited by students in Alex's university-level teaching experience provided one source of judgement, although he resorted to them only when (as with essay #4) the rating criteria he normally employed did not yield a satisfactory judgement:

Now, OK, so, it's not right on task. She's not really right on task, but it has to be an average, a 3 or a 4. Let's put, let me put down a 4-. OK. It may go down to a 3. OK, uh, I mean although the content, most of the content is not relevant to the task, she has been able to display, hmm, uh, some control of the language. And uh, and there are some words like "elegant," "appreciate," uh, hmm words like that, some strange words, but uh, this has to be uh, an average in this bunch of scripts.

The result of Alex's tendency to establish criteria from comparing papers to each other, as well as to norms developed by observing the writings of his university-level students, is shown in the predominance of not only interpretive but also judgemental comments with a self-monitoring focus throughout his rating session (Figures 13 and 14), but particularly in the final phase where Alex confirmed his ratings (Figures 17 and 18). As for the relative importance of language versus rhetoric and ideas, both during the initial read-through of all papers, in the second phase of his rating session (see Figures 15 and 16), and in the final reading (see Figures 17 and 18) Alex's interpretive comments showed a slight preference for language while his judgemental comments showed a more marked preference for ideas, rhetoric and organization.
Figure 13: Proportional frequencies of interpretive strategies during Alex's rating session.

Figure 14: Proportional frequencies of judgemental strategies during Alex's rating session.
Figure 15: Proportional frequencies of interpretive strategies during phase II of Alex's rating session.

Figure 16: Proportional frequencies of judgemental strategies during phase II of Alex's rating session.
Figure 17: Proportional frequencies of interpretive strategies during phase III of Alex's rating session.

Figure 18: Proportional frequencies of judgemental strategies during phase III of Alex's rating session.
Although Alex was thus balanced in assessing both language and ideas and rhetoric, his protocols revealed that the only criteria he had used with any frequency in scoring papers were "reasoning, logic and topic development," "task fulfilment," "ideas and rhetoric overall" and "language overall." The first three qualities were weighed to provide the first approximation of the score, which was subsequently refined by assessing language control. Alex felt that "below average" essays provided, at best, a minimal, and possibly incomplete, answer to the question posed in the prompt; among these, he awarded a 1 to essays deficient in language control and organization, and a 2 to those which showed some control of language and/or some awareness of basic structural requirements. Essays Alex judged to be "average" provided either incomplete or one-sided arguments; more specifically, he awarded a 3 to essays displaying limited task fulfilment, basic grammatical errors, or logical confusion, and a 4 to essays that largely fulfilled the task but suffered from gross errors in logic, irrelevant arguments, or inadequate development. Finally he judged those essays to be "above average" which amply fulfilled the task, besides being fluent and creative. Significantly, the language control of some essays to which Alex awarded a 5 was lower than what was displayed by essays he awarded a 4. However, he gave a 6 only to essays which not only fulfilled all aspects of the task, but were also free of all but minor linguistic errors.

Significantly, although Alex was familiar with numerous rating scales, he felt that he had been little influenced by them in setting up rating criteria, beyond learning to pay attention to such broad elements of communicative competence as language use or organization:

"If I had never been in language teaching, I would have never learnt to look at a piece of writing from at least three primary traits [identified earlier as content, language and organization]. If I were just somebody, you know, teaching geography and somebody gave me a piece of writing and asked me "So how would you assess this?" I would say: "Oh, I don't know." So how am I going to say that the rating scale doesn't affect me, or affect me to what extent?"

Consequently, assessment experience, particularly the extensive feedback Alex regularly received from the examination authority he worked for in his native country regarding his consistency, severity and ability to discriminate among candidates' proficiency levels, exerted
only a general influence on Alex's establishment of standards in the rating session for the present study.

Instead of relying on assessment experience to construct his criteria, then, Alex started with the assumption that TOEFL candidates were applying for admission to a North American university. Such knowledge he used, first of all, to establish norms: since he had been teaching first and second-year university students in his native country, he made the assumption that these students resembled the candidate pool for the TOEFL, and could provide reference points the essays he was rating. However, the contrasts that Alex drew between university students and secondary school students also helped him to construct a developmental trajectory for language learners, and helped to further specify his expectations:

As I was reading [THE PROTOCOL] just now, my question was, "How much I should credit a candidate who failed to complete the task but at the same time has been able to display a level of language control?" Because in university assessment, at least in the program that I used to teach in, we would actually give very low marks, sometimes even a 1, for failing to complete the task, even though the language is, is fine. [...] Well academic task, at least you have a role to assume, you have to know who you write to, you have to know why you are writing it and you have to realize these in the writing. [these skills] are also expected of secondary school students but that expectation is usually not realistic in the sense that you more or less bother with syntax and word choice and word formation problems more than audience analysis or audience orientations and genre.

Alex also combined his assumptions concerning the test taker population with his perception of the purpose of the TOEFL to generate additional rating criteria related to task fulfilment:

Because I know this is TOEFL, and because I know that the task requirement is not very specific, it's more like "you have a task because you want to give them the, some, some context to write something." So, those aspects are not, I felt, at one point, not very important. So, I would still try to give some of these candidates a 3 or a 2, depending on their display of language. I probably may have given one a 4, knowing that he didn't complete the tasks but still displayed a certain level of proficiency.

In addition, relying on his familiarity with both "Western" and "Oriental" cultural traits, constructs that he employed in the interview, he could even discern different trajectories of language learning, even if his own experience suggested a fairly strong emphasis on grammar:

I understand that this is in a different context, especially in terms of culture, it's, it's, it's different types of learners, and, yeah you are right; we all come in with our own beliefs and biases. ... But I guess what happens is, and this is the point that you make, that I come with biases, because my experience is more restricted to one country more or less. And, uh, I think some people acquire English in a different way. I mean they didn't-. For example some people
learn English, because they are immersed, I mean in immersion. They are in an English speaking environment. Not that they have to write a lot of English, so their, their spelling may be terrible, I mean they are just imitating sounds. But they can have all the fluency. Like what I said here on one of the candidates. "He must be a very fluent speaker because you can tell from the flow of the language, but the spelling, you know, it's terrible." So you know that that particular person may have learnt English, uh, not in the, not in its written format. ... Whereas in (NATIVE COUNTRY) it's not like that. (WE PLACE A) fairly strong emphasis on form, yes. That's true. Not only that, weak in terms of making new points in content.

At the same time, while he placed a premium on language control, Alex regarded it merely as a necessary but insufficient condition for a high score. He didn't see language control as a sign of intelligence, although he did regard its absence as a sign of insufficient learning and, in cases of errors that he had seldom seen at the university level, he could also conclude that a candidate was not ready for university:

Word, sentence level, structure, common errors by learners (WHO SPEAK ALEX'S L1), and also against the advice they usually get from teachers, you know. Uh, that's why I don't expect to see university students making mistakes like "I suggest you to go" when I know that this is something that teachers advise against in secondary school days. So only a student from (MY COUNTRY) who, who hasn't learnt that language effectively would make that sort of mistakes. You see what I mean? I mean not just acquisition, but also the pedagogic focuses. And if a learner, well, you know, makes this type of errors at university level, you can tell that he didn't learn the language very well in, effectively, in secondary school, not that the teacher didn't teach, it's in the syllabus somewhere. It's just that the error is so stubborn, you know, or he wasn't paying attention. So it's not just this, this stage-wise, uh, natural route of learning principle, but it's also looking at learning language at the macro-level, the syllabus.

To get a high score from Alex, then, a candidate had to demonstrate other skills which to him were indicative of a higher level of proficiency. One such skill was fluency, which Alex took to be a sign of progress from sentence-level mastery to discourse-level mastery. Another was a movement from formulaic organizational patterns to the achievement of overall coherence and cohesion. Finally, to complete his trajectory, Alex only had to define the upper end, and he used his top ESL students at the university in his native country as his reference group, combined with his perception of the competence of native speakers:

Well, I guess my knowledge [is] that even fairly educated native speakers can make errors. OK. That knowledge informs me that occasional grammar errors are no obstacles to giving a person, especially an ESL learner, a top mark. Content is important and if he has a fairly good organization, if the idea is well formulated, you know, I understand it, making a point and if that point is novel and relevant, a 6. I understand that there is no way ... to compare the proficiency of these learners to educated native speakers. It's more like, "that ceiling [represented by a 6] is there for people who have learnt only that many years of ESL," and [candidates getting a 6 here perform like] the kind of students that I see at universities that are getting the best English grades.
Having constructed a model of language development, Alex refined it further with the aid of his experience in teaching writing in the sciences and his knowledge of the pedagogical and theoretical literature:

This is my definition actually: "whatever you have to do, to write in school." So lab reports would fall into that category. And all the research articles, or terms papers you write for your prof, those are all academic writings. And I feel that these are more like academic writings because you are asked to, you are doing the same sort of rhetorical functions, you group and organize information, you compare, you contrast, you argue and you try to give reasons for your statements, these are academic. [...] it's in the literature, in like, uh, I mean like English for Science and Technology by Trimble (1985) these sort of books. Oh yeah, and Genre Analysis, uh, by John Swales (1990), there is a famous book on analyzing research articles. How various rhetorical functions are realized in writing research articles. How you produce an abstract, how you play the role in arguing for or against something, how you compare, how you illustrate, diagrams, technical writings. You know, to me these are all skills required by students in their pursuit for academic qualifications. So to me that notion is an English for Special Purposes, you know in an academic setting.

Cultural norms were another important influence, and they entered Alex's rating scheme through an awareness of different pedagogical practices, discussed above, as well as an awareness of contrasting attitudes to original thought:

As I said, plagiarism is a cultural thing. For many ("ORIENTAL") learners, in their mind, to speak in the language of somebody else is only the right thing to do. You don't speak what you speak, you speak what the sages speak! ... I think there are some researchers looking into the question of plagiarism, and think this is probably a notion that is more relevant to Western culture than to Eastern culture, because in the West you do encourage, you know, novel thinking, creation, whereas in the East it's a different philosophy, you see? So, I guess, if you ask me out of those three aspects, language, content, and organization, so I would associate language ability more with language control and organization, because those are the things that are teachable and learnable.

Consequently, Alex was willing to accept pedestrian content and a lack of commitment to the ideas expressed in students' compositions, in keeping with the "Oriental" approach to argumentation that, as an examiner, he had frequently observed.

By contrast, Alex was much less forgiving of arguments that were insufficiently developed or supported, especially if they involved sweeping generalizations. This attitude was in keeping with his emphasis on organization and language control, although, interestingly, he ascribed it, rather, to his acceptance of certain "Western" cultural values:

I think [DISLIKE OF SWEEPING GENERALIZATIONS] has to do with my Western education, that I think claims should be supported either by examples or by reasoning, or by conceptual links, you know. I do believe, I mean, I think that an unsupported claim is worse than silence, you know. [...] In writing you are out there to communicate. And what is communication? The problem I have with some of my engineering students is that they always think that communication is about information transfer, more like transmitting information. I said, "Look,
this world is full of information," alright? ... And, then, if you are talking about transmitting information, you are not communicating. You are making a point ... and when you make a point, you support it.

Seen in this light, Alex's emphasis on language control was not merely a product of teaching practice or of ideas about learners' developmental trajectory, but also an outcome of cultural attitudes, reinforced by Alex's experience, as a rater, of how "Oriental" candidates handled culturally inappropriate essay prompts:

To many writers, Oriental writers, [SOME ESSAY TOPICS ARE] not something that they can put themselves into, not something that they commit themselves to ... To them there is no such question (AS THE ONE POSED BY ONE OF THE ESSAY PROMPTS USED IN THE RESEARCH PROJECT)! You know, this question doesn't exist! So, he is playing the game, you know what I mean? The game. So what do you [look for]? You don't look for in a certain ways, at least from my view, you don't look for personal commitment or position. You look at the display of language. [...] I mean, I mean we, we might as well cherish personal knowledge, personal positions, and writings that are committed. We can encourage that, we should, in, in classroom. But, again, we have to see that this is a notion that is perhaps more relevant to the West than to the East. So these are my feelings towards the inter-cultural aspect of a public exam like this.

Unfortunately, this clash of cultural values begs the question of how Alex decided on accepting or rejecting "Western" values, a question I cannot adequately address here. I can merely reiterate that cultural background, like everything else, is individually constructed and it is impossible to infer from a knowledge of ethnic origins how individuals absorb different influences. This, of course, is as true of writers as of raters, and provides solid justification for keeping the ethnic and linguistic identities of candidates confidential during public examinations.

6.5 Summary

Alex cited not only his wide-ranging teaching experiences, but also the findings of researchers (such as Pienemann, Johnston and Brindley (1988)) in support of the notion that writers went through identifiable morpho-syntactic stages in their development. He even extended, albeit informally, the Teachability Hypothesis to cover discourse-level and sociolinguistic competence. Although he had extensive assessment experience, Alex continued to rely primarily on his teaching experiences for his construction of a developmental trajectory, combined occasionally with textbook prescriptions. However, thanks to his assessment experience he
was able to combine his inferences from teaching with numerous specific expectations concerning test takers and test use. Above all, while he was aware of the need for flexibility in expectations concerning task fulfilment, linguistic accuracy, and creativity, given the diverse cultural backgrounds of writers, he was able to cite the specific use made of TOEFL scores (to screen candidates for admission to North American universities based on their English proficiency) in support of deriving a uniform set of rating criteria for the compositions he was assessing.
Chapter 7: Jo

Jo was a graduate student and a NS, in his 40s, who had extensive experience not only with teaching ESL, but also with teacher training, test development, and proficiency testing. I chose him for the present study partly because of the extent of his experience in both teaching and assessing ESL proficiency, and partly because his concurrent verbal protocols had revealed that he had used a particularly elaborate strategy in assessing the compositions for the Cumming et al. (1998) study.

7.1 Research question A: In what principled way did the raters of ESL compositions participating in the study relate the performance embodied in a written text to a writer's level of proficiency, and how did background factors influence this process?

In his interview, if not in his concurrent verbal protocols, Jo made frequent associations between a writer's performance on a TOEFL essay and her level of proficiency. He attributed certain textual features, such as the ability to use a range of complex sentence types, to advanced writers and other features, such as irrelevant content, to less advanced writers. In explaining this tendency to associate proficiency level with performances through an informal developmental trajectory, Jo relied on a wide range of scoring rubrics that he had worked with, and assumed these to be reflective of a writer's development (cf. Brindley, 1998):

[LANGUAGE BENCHMARKS] are, you know, a continuum, which, which is a norm-referenced approach because it ranks people, but each benchmark is described in terms of what the learner can do, what kind of structures they can produce, and the level of difficulty. So, and also I did some marking with the TWE scales [EDUCATIONAL TESTING SERVICE, 1989], also, and those have, well there were six levels, but they were described, you know, in terms of increasing levels of ability, although they were all much higher so, in a way there was less distinction in that scale than there was in the one I was trying to create here. Uh, so, I guess, yeah, I guess, that helped me to justify, too, what I was doing because I would feel otherwise "on what basis am I saying that this is better, you know, than that." And the other thing I found was that length influenced me to some extent, too, and I had to get around the fact that somebody may have written something lengthy, Uh, I think it comes from having done the placement testing when I worked in [AN INTENSIVE ENGLISH PROGRAM AT A NORTH AMERICAN UNIVERSITY], because one of the things we, we had a speeded test and one of the things that we would notice is "Look how much this person wrote in this short time." So the initial reaction was "must be a more advanced learner," and then you would look more closely and maybe you would end up refuting that, but, but that would be a first assumption and I guess I am still influenced by that.

89
Note, however, having derived broad rating criteria from existing scales, Jo did not merely apply his ideas of a developmental trajectory, but refined his assessments by comparing compositions with each other within the corpus he had been asked to evaluate for the Cumming et al. (1998) study.

In addition to estimating proficiency from performance, Jo was one of two raters (Sam being the other) to also assess a writer's potential in cases where he could clearly identify the negative effects of certain situational factors. Nor did Jo limit such compensation to borderline "pass/fail" cases, as the following protocol, concerning an unfinished composition (#95), shows:

Looks like [THE AUTHOR OF ESSAY #95] had a time problem here. [...] Uh, I mean, I give it a 1, because it doesn't give any argument. And yet I'm tempted to give it a 2, because I feel that the person has done only an introduction and given more time, and I hate speeded tests, so, given more time, I feel that this person might have been able to organize an essay that would be worthy of a 2. [...] I'm giving it a 2, because it looks like just the beginning of a decent essay for someone who might be capable of a decent essay, let's see, or perhaps a barely acceptable essay, for someone who ran out of time. So I'm giving it a 2.

Jo's attitude towards compensating for situational factors was expressed clearly in the questionnaire, where he stated that he was aware of the influence of both time constraints and topic effects on performance. He elaborated on this during the interview by mentioning his awareness of the negative impact of pressure on candidates writing gate-keeping tests, which, in turn, puts pressure on the raters themselves:

I am always trying to be so fair, and I think, "Well, what does this mean to the learner and what can they maybe not do if, if they don't get a certain mark in this test? What is it preventing them from doing in their lives," because sometimes it can be a really important thing. This wasn't, I mean it was a practice thing, but I came to, I tried to come to it with the same energy. Partly because I know what [THE TWE] can do to people, too, so, maybe university entrance or whatever.

7.2 Research question B: How did raters of ESL compositions interpret the role of writing prompts in performance-based assessment, and how did background factors influence this process?

Being keenly conscious of the impact of prompts on performance Jo was, in principle, willing to compensate writers for underachievement attributable to a poor prompt. He certainly took topic effects into account by deriving his rating criteria at least partly from analyzing the
textual qualities displayed by the corpus he was rating. On the other hand, Jo assumed that the writing prompts for such a widely used test as the TOEFL would have already been vetted by test developers. Consequently, while at the beginning of the rating session he analyzed the essay prompts from the point of view of accessibility to candidates, Jo's principal concern was actually with the level of challenge, as well as level of difficulty, posed by the writing prompts:

So, in beginning to rate I guess, I already start to think about the degree to which the prompt challenges the learner and uh, I haven't looked at the essays yet but I wonder whether, uh, whether that might influence me, whether I might uh I might think uh, think when I'm marking a mediocre essay that "well it was on a really difficult topic" or "it was harder to address that topic" or when I write, when I mark a better uh, essay on a, on a more uh, what would seem a more basic or straightforward topic uh, would I, would I take that into consideration. I'm not sure.

This attitude was reiterat in a comment Jo made in his interview, which also illustrates his extreme reluctance to actually compensate writers for adverse topics even if he recognized the latter as impediments:

[DID I COMPENSATE A WRITER] in response to a particular prompt that I didn't like? Because that would be one reason; if there was a prompt that I thought didn't elicit, you know, a sufficient response, I might have thought that they could have done better if they had written on another prompt. Why would I do that? Why would I assume they had more? ... It may come from the belief that people are never able to show their true potential under exam circumstances, but I still don't think it is, you know? I mean, my philosophy would be that it is not the role of the assessor to second-guess that so I don't know why I did it in those circumstances.

In the case of the rating session Jo undertook for the Cumming et al. (1998) study, he actually concluded early on that since certain candidates (in this case, the authors of essays #1 and #88) had managed to write well on both limiting and challenging topics, there was little point in analyzing topic effects:

Just an aside here but I like the fact that once one gets scoring these papers, the topic kind of fades into the background. I'm no longer looking to see what the topic is, and what exactly was asked in the prompt because I have a sense now from having seen the exemplars uh, what, what the learners are capable of and I'm sort of trying to just put that into the back of my mind so it's not lost. I still have a sense of what the prompt is but I'm not letting that totally uh, cloud my, my reading of the essay or, or become a huge factor in, in the scoring.

Jo's general attitude towards writing prompts had also been influenced by his experiences with tests and test development, which had shown him that writers reacted to prompts in unpredictable ways. In consequence, as long as candidates were not merely avoiding the topic
in order to recite memorized bits of language, Jo felt that they should be given considerable latitude. As Jo confirmed in his interview, for him the purpose of a prompt was:

> to bring forth the best possible sample because I think that's all the prompt is ... something that is supposed to motivate [learners] to give their best performance. So first I have to examine the prompt to see whether it does indeed do that because if it disadvantages the learner, I have to know that, I have to know that that prompt didn't ask the person to elevate themselves to this level and maybe they could have. So, I guess I am doing that and once I have satisfied myself that the prompt hasn't disadvantaged them, then I take it just as a piece of communication, so I can leave it be. So, it's only relevant to the extent that it may have influenced the performance in a negative way.

Given that Jo thus regarded prompts as vehicles for eliciting a performance, it is not surprising that he evaluated task fulfilment only 8 times, representing 0.4% of the total number of coded behaviours in his protocols. Although he commented on the specific demands of prompts more frequently (20 times, representing 1% of all his coded behaviours), he referred to such demands, like the assessment of task fulfilment, usually when writers hadn't fulfilled them (e.g., the writer of #143, who expressed a preference instead of stating her agreement or disagreement with a point of view, as the prompt had demanded).

In addition to regarding prompts as vehicles for eliciting a performance, Jo resolutely avoided any speculation concerning writers' ethno-linguistic or socio-economic backgrounds, even when such information was offered by the writers themselves. In explaining such detachment, Jo referred to the subjective nature of performance-based assessment, which, he felt, it was best not to compound by introducing additional variables, such as information about writers' backgrounds. Indeed, Jo favoured anonymous assessments even in classroom situations, unless and until he became equally familiar with all his students, an unlikely situation in his case given that his recent teaching experience had mostly involved very large classes:

> I know when I mark class assignments I don't look at a person's name. In fact if I know what class [THEY BELONG to] at all, I don't look to see who is writing the papers because I don't want to know. Why is that? I guess it's just [...] uh, I guess it's the need to let the information I have speak for itself, without imposing any other ideas. I have no idea where that would come from. [...] When I teach a non-credit ESL writing class in that I get to know the students, I bond with them, I see whether they're improving I know their systematic errors. But when there are stakes involved, or credit, I don't want, I don't want that subjectivity. I guess it is, it is some deep-seated belief that to be fair. Yeah. If I pick up a paper written by a student I know and then I pick one up written by a name I don't recognize then the other person is being treated as an unknown entity. So I treat them all as unknown entities unless I know them all.
At the same time, Jo did make specific assumptions about the candidates who wrote the essays he was rating. Some of these assumptions came from his knowledge about test use; for example, that the TOEFL writing test was usually taken by university applicants. Other assumptions arose from Jo’s practical experience with assessment, which told him, for example, that candidates used a range of strategies to maximize their score, and that these had to be identified if a writer's proficiency was to be estimated:

I can see here that I wanted [writers] to address the topic and I tried, I tried not to be too influenced by the fact that they didn't address it, because, I mean, maybe they read it and misunderstood it, but on the other hand I thought it could be avoidance, they may have made the topic a little bit more simple for themselves, or in that one case, there was that one paper where it looked like the person had almost brought in pre-planned vocabulary and, and was structuring the topic to fit this vocabulary. So I guess I try to look for tricks, too, for things I thought they could do to, uh, maybe influence a marker. "Maybe I can't write on this topic but if I write on that topic the marker may take, have mercy on me."

Finally, while Jo regarded the candidates writing the TOEFL essay for the present research project as forming an essentially homogeneous group, characterized as "university applicants," and made no distinctions among them based on ethnic or socio-economic grounds, he was aware of the existence of personality differences among writers, and of the need to factor these into his assessments:

Part of strong writing, I think, is to be able to support your ideas, but if you can write with good structure and you have boring ideas, I don't think you should be penalized, you know. So, I was trying also not to do that either. That comes from teaching. Yeah, that comes from being in the classroom and knowing the personalities of different learners, and realizing that not everybody brings the same energy and the same background experience to the task and yet, uh, they, they can be competent in what they're doing, you know. And I think that this happens in all kinds of assessment, especially listening-speaking, that personality can influence an assessor to a great extent, you know an outgoingness. So I try not to read that outgoingness in the writing as skill or competence.

7.3 Research question C: What reading strategies did raters of ESL compositions establish (to deal with both individual compositions and a corpus of compositions) and how did background factors influence these?

The scoring procedure Jo followed did not fit neatly into any of the categories established in Milanovic, et al.'s (1995) study of raters' behaviour, which envisaged a maximum of two readings for each essay. Jo's protocols actually revealed a four-stage procedure, involving multiple readings of essays, each time with a different purpose. In the first stage of his
reading of the corpus, Jo simply sorted a few compositions into five piles to establish anchors for his scoring scale:

#62. OK, so, here we're certainly getting to the bottom of the scale here, somewhere low. [QUOTE] So again, now they haven't even addressed the topic [...] I think there's a misunderstanding of the prompt here. [QUOTE] So I'm going to say that's a low. That is a low one, not necessarily a 1 but a low one, one being one of many.

The end result of the first stage was that Jo established one pile each for the end-points, one for the middle point, and one each for points halfway between the middle and the ends:

So now, I sort of have these five piles. I have like a really low pile and a really middle pile and a really high pile and then I have two piles that kind of fall in between, so like a low-mid and a mid-high and probably what I'll do is I'll go through the whole pile of essays and I'll allocate them to those five piles and then once I've done that, I'll look at, I'll look at the highest pile to decide whether those are 5s or 6s and the second highest to decide whether they're maybe 4s or 5s and the middle pile to decide whether they're uh 2, 3 or 4, etc.

Subsequently, Jo scanned a few more compositions and re-sorted them into six piles from the original five; at this stage the end-points were clearly defined in his mind and he had exemplars for each of the other levels, as well. Following this, in the third stage, Jo read through all the papers, including the ones he had already seen, and assigned them actual scores:

The next one I pick up is number 62 and I have to ask myself, OK is, is this... the dividing point between a 1 and a 2? It's definitely longer than the other ones that I've given 1. It's three paragraphs long and they're three nicely formatted paragraphs. The first has a, has an introductory sentence, "THE LOCATION OF..." So, this, this again is disorganized but somehow there's been an attempt here to develop the ideas. So, it seems to me that, that the 1s were incapable of developing an idea comprehensively and this 2 is, is one in which the writer has made a, a valiant attempt to organize advantages and disadvantages ... but has had difficulty with structure. Actually not many difficulties with spelling as I can see here. Uh, ... but difficulty with structure and, I think a little bit of difficulty with building. You know, building one idea on, on another in a, in a logical sequence but still I'll --- I'm going to give this a 2...

During this stage Jo also established a rating scale, and frequently compared compositions to exemplars. Papers to which he could not immediately assign a score since they didn't fit neatly into any of the six emerging piles he set aside for a second reading. Finally, once he had assigned a tentative grade to all the compositions, and had fully established his rating scheme, Jo took a final pass through the corpus and confirmed the scores he had assigned, as in the following example:

Number 62. I'll put that in the clear 2 pile.
In spite of the complexity of the strategy, Jo was able to rationalize his actions throughout the protocols; 1.7% of all his coded behaviours actually consisted of "deciding on micro-strategy for reading or rating" (JS1 — see Table 1). In turn, he ascribed his ability for introspection to assessment experience. The key aspect of his strategy was taking multiple passes and this he related, first of all, to the way he structured his feedback to learners in his writing classes:

I am trying to look at structure, or I look at the overall ability to communicate meaning, before I look at smaller errors that might not affect the overall message, uh, like, punctuation or spelling. ... I think I probably mark that way when I teach writing. So I go, uh, I may make a few passes through an essay, I may give it back to a learner three times to improve, and so the first time I look at those overall structural things, before I start really nitpicking, because I don't want them to get, uh, too discouraged, so I don't point out all the errors the first time through, and then I, and then I get more and more specific. So I think, I was thinking of that ... marking with that process in mind. To me the first step was more important than the subsequent. So, I didn't realize that but that's exactly what I was doing.

Another contributing factor was Jo's extensive assessment experience, which had taught him that different readings of the same text invariably produced different impressions, that it was impossible to sustain a high level of energy during a lengthy rating session, that there were inevitable grey areas between bands on any rating scale, and that decisions made by assessors could have a serious impact on candidates' lives. In addition, while Jo's rating experience usually involved teams of markers, who would periodically reconcile their ratings, this was an individual assignment, so Jo felt the need to read papers several times in order to compensate for the absence of external checks. Finally, Jo also referred to his understanding of the literature on performance-based assessment to support his policy of multiple readings.

Related to the policy of multiple readings was Jo's use of interim assessments, which he verbalized not only between but also within readings, as the following protocol demonstrates:

I'm just reading silently paper number 8... The person leads into the topic with a general statement ... and then narrows it down to what these rooms are ... mentions that each of them has a purpose ... then narrows the topic further to a special room in each house that is considered the most special uh, and then indicates that that is the family room. So that, that's a kind of introduction I sort of like ... Now I'm just looking at ... some of the analytical features of this piece of writing and it doesn't look very sophisticated. ... So this person has some sense of how to organize but is having difficulty with uh, putting that, uh, into practice using English structure and grammar. So already now I'm thinking it's in the 3/4 range because it's not a 1/2. The person isn't struggling with comprehensibility and it's not a 5/6 in that it's not particularly sophisticated ...
In the interview Jo explained his practice of giving interim assessments with reference to his readings concerning the logical positivist position on scientific explanation, which focused heavily on hypothesis-testing as a strategy; indeed, he had found this a useful strategy not only in scoring compositions, but also in teaching reading comprehension to ESL students.

The other outstanding feature of Jo's rating procedure was the construction of descriptors based on the qualities of the corpus itself, a behaviour attributed to experience with test development:

I think I let the material tell me what the descriptors were after I ranked it. ... Where that comes from? Ah, test development. That's where that comes from because when we, some of the tests that I have developed, we made tasks, had learners trial them, brought back the tasks to see what the learners brought to them, knowing the learners' levels and then establish criteria based on what the learners did. So that's that. It's a backwards way of doing it but that's sort of what I was doing here, not having criteria, developing a set of criteria through responses to tasks.

Finally, while Jo was aware of the specific circumstances surrounding the rating task for the present study, particularly the absence of an a priori rating scale, he stated in his interview that he would have considered the qualities of the corpus even when applying a predetermined rating scale. This further underscores the problems that experienced raters have had with the mechanical application of rating scale descriptors, a problem which compelled Jo to use exemplars, another outstanding feature of his strategy:

First it would be really important for me to know who else is using [THE DESCRIPTORS] and to sit with them and discuss our interpretation of some of the terms, especially adjectives, terms that are questionable. And, then to look at a sample of papers, probably 15-20 papers, and double-mark them and look at what we are doing and decide as a group and then if there is a discrepancy to decide, "well, here, you seem to think this is sophisticated and I don't; what, what do you think is 'sophisticated'?" and then to point out areas within the paper. So that, that would be really important to me before I would use any scale would be, you know, have some training in how to apply it and, and, then the whole idea of discovering through the exemplars what some of the criteria are is, I guess a way of, rather than coming with a preconceived notion of saying that "OK, a good paper will have this or this" and then looking through them and saying "well none of them have it, so, there isn't a single 6 in there," I would prefer to let the papers speak to me, because maybe people, sometimes the learners know more about how to address the task.

7.4 Research question D: By what process did raters of ESL compositions construct specific rating criteria, and how did background factors influence this process?

Jo started by establishing an interim, 3-point scoring scheme for the compositions, and expanded it first to a 5-point scheme and then to a 6-point scheme. In his interview he
justified adopting this procedure on the grounds that a 3-point scheme provided a useful first approximation:

You know, I mean that is sort of the simplest way, I guess, is to approach the range, to see what the two extremes are and what falls in the middle ... And it's much easier to make a binary distinction among these three piles, so that's what I was thinking of, mathematically: "let's divide it into these three piles and then let's divide the piles and that will simplify it, rather than dividing this whole mass into 6 piles.

Jo described this interim scoring scheme in his concurrent verbal protocols, at the point when he began to assign grades to the compositions he had until then been sorting into piles.

So far what I'm thinking in terms of my ranking is that the 1s and 2s have the comprehensibility problems. The 3s and 4s are struggling with, probably with the structure and the presentation and the 5s and 6s they have it, but, they have everything that they need. They have organization. They have the structure. They have uh, they have, the, the grammatical competence but the 5s are just lacking. They're lacking the, the fluency I think and the sophistication that the 6s have.

Jo’s interim rating scale shows already that he weighed both language and ideas and rhetoric in assessing compositions. This is reflected in the distribution of coded behaviours throughout his protocols. Overall (Figures 19 and 20) both Jo’s interpretive and judgemental comments were weighed heavily towards self-monitoring behaviours, and this reflected his frequent use of quotations from texts and also his tendency to compare compositions within the corpus, articulate and justify scoring decisions, and articulate rating strategies. The rest of his interpretive comments were somewhat biased towards a language focus, while his judgemental comments were almost evenly split between rhetorical-ideational and language foci. The same pattern was maintained throughout the various stages of his protocols (Figures 21 to 26).

Although Jo did not articulate the elements of the expanded 6-point scale, he frequently defined his individual scale points, starting with the criteria for getting either the top or bottom score. He felt that the authors of essays he scored at the lowest level (e.g., #84) were incapable of developing an idea comprehensibly or responding even minimally to the demands of the prompt. The authors of essays Jo judged to be 2s (e.g., #48) were at least able to present coherent arguments in favour of their stances, even if they couldn’t develop their ideas fully and committed serious linguistic and/or organizational errors. Although Jo’s crite-
Figure 19: Proportional frequencies of interpretive strategies during Jo’s rating session.

Figure 20: Proportional frequencies of judgemental strategies during Jo’s rating session.
Figure 21: Proportional frequencies of interpretive strategies during phase II of Jo's rating session.

Figure 22: Proportional frequencies of judgemental strategies during phase II of Jo's rating session.
Figure 23: Proportional frequencies of interpretive strategies during phase III of Jo's rating session.

Figure 24: Proportional frequencies of judgemental strategies during phase III of Jo's rating session.
Figure 25: Proportional frequencies of interpretive strategies during phase IV of Jo's rating session.

Figure 26: Proportional frequencies of judgemental strategies during phase IV of Jo's rating session.
ria for getting scores of 3 or 4 were more elusive, he did define them in the third stage of the protocols. With a few exceptions he scored those essays 3 which he thought contained the basic structural elements but were either insufficiently developed, or error-prone (e.g., #63). By contrast, essays which he thought provided well-developed answers and displayed some mastery of complex structure but were devoid of creativity or sophistication in both ideas and language (e.g., #111) he scored 4.

The key criteria in Jo’s rating scheme, therefore, were "comprehensibility," which he used to separate the lower level papers from the middling ones, and "sophistication," which he used to distinguish papers at the upper levels from those in the middle. Both of them Jo identified at the outset of the interview as the ends of the continuum of language learning, and they enabled him to make initial distinctions between papers in a manner recalling the binary decision-making scheme advocated by Upshur and Turner (1995):

So this person [the author of essay #8] has some sense of how to organize but is having difficulty with uh, putting that in, putting that into uh, into practice using English structure and grammar. So already now I'm thinking it's in the 3/4 range because it's, it's not a 1/2; the person isn't struggling with comprehensibility. And, it's not a 5/6, in that it's not particularly sophisticated.

In explaining his use of comprehensibility as a fundamental criterion in his interview, Jo showed the influence of communicative language teaching:

So that comes, I guess, from, just basic notions about communication. I mean, the idea that language is a medium of communication and that it's, the point is to get your ideas across. So are you doing that? Are you succeeding? So, I guess I am looking at some primary objective, which would be communicate your message and then some secondary objectives which are, have to do, with how you manage to communicate it?

Such a focus on comprehensibility may be related partly to Jo's adoption of the distinction between global errors and local errors (cf. Lane & Lange, 1995) in applying a process approach to writing instruction, where successive episodes of feedback were meant to address errors interfering with communication to differing degrees. However, the communicative approach also represents, fundamentally, an assumption about language learning; in this case, Jo maintained that it was in the effort to communicate that mastery of structures would be achieved. Therefore, it made sense for Jo to use "comprehensibility" as a minimum requirement, and he used his assessment experience to come up with a measure for it:
How many times do I have to read it before I say "Aha." ... Probably three times through, would be. I would be concerned, on the third reading I would be noting in my mind "don't forget this is your third reading." Because I know that if you do read something over and over again, the danger is that you are going to, on the 3rd or 4th time, think it's pretty good, because you figured out what it means, which means you filled in those little gaps. So, you have to be always aware of that.

"Sophistication," at the other end of the spectrum, Jo defined at times as the mastery of vocabulary, whose appreciation stemmed from his reading experiences:

Sophistication. It's, it's level of vocabulary, it's turn of a phrase, like language use, the ability to put a phrase together, so that it sounds either native-like or, despite the fact that it may not be exactly what a native speaker would say, it's really creative and maybe even better, maybe it turns the phrase in an even better way. ... [its use as a criterion comes] probably from things like, really loving things like Shakespeare. I love the use of the language and the way the words are put together, and I, I find that a really beautiful thing.

It is important to note, however, that while Jo viewed sophistication as the ultimate goal of language learners, he did not equate this goal with native-like competence; instead, in a rare example of voicing cultural values, he wanted learners to preserve their distinct identities:

[I have] not just an acceptance of, but almost a protective feeling toward a person's own voice, the person's accent, uh, presentation, I don't like to see that lost and that might come from my artistic background. I think that's a real shame, if we homogenize that, and we lose the authenticity of that voice. ... I guess having been around a multicultural mix I would kind of worry if everybody sounded exactly the same.

Nevertheless, given his definition of writers' distinct identities, Jo certainly expects writers at the top level to produce essays that were essentially error-free, as his assessment of essay #128 shows:

Number 128[.QUOTE] Uh. I like the opening, even though it doesn't address the topic[.QUOTE]. So already I'm thinking this is a 6. For-, because it's flowing, it's uh well expressed, it's error free so far [...]. I'll just read this silently, but I think it's a 6. ... Absolutely. They are using words [.QUOTE] and, so far error-free in terms of structure and grammar [.QUOTE]. Use of gerund as a noun. Oh, misspelling... Missing article. ... Using "me" as a subject, instead of "I". Uh, I'm thinking 5 now. It's not that I expect the 6 to be error-free. It's just that this I started out impressive, and then became less so.[...]. Uh. So just in making the final comparison I decided to give it a 5, because there are enough awkward statements in it. I mean, it seems to flow, and it seems to be coherent, but there are enough awkward statements, such as uh "I THINK GIVING...". Things like that. It's just - - -. It's sophisticated, but there are little bits of awkwardness in it, that just cause me to give it a 5, rather than a 6.

Sophistication also emerged for Jo in the realm of ideas, through writer's ability to take an abstract view, as well as follow the conventions of academic writing, which, as his comments on genre suggested, included abstraction, detachment and a formal tone rather than personalized and experiential writing. These are qualities that Jo, under the influence of rating scales, took to represent a higher level of cognitive ability and regarded as the hallmarks of academic
writing. Consequently, since these abilities were essential for success at university, and since, in Jo's view, the majority of candidates for the TOEFL were prospective university students, "sophistication" naturally became the key requirement for Jo's placement of compositions in the upper two bands.

Even Jo's definition of the middle levels of proficiency reflected ideas of progress: adherence to the traditional system of essay and paragraph organization usually sufficed to merit a 3 or a 4 in his scheme, as long as it was accompanied by some mastery of forms, because it was a sound strategy for lower-level writers, to be discarded in subsequent stages. Such judgements were the products of both the way Jo had learnt to write himself, and of his subsequent teaching experience and exposure to writing textbooks. It is important to note, however, that not all of his rating criteria arose out of Jo's assumptions concerning language development. For example, Jo's stress on the mastery of formal aspects of English writing, while embedded in the concepts of comprehensibility and sophistication, also reflected his experiences with teaching grammar. In addition, through his work on test development, Jo had become aware of just how harshly people outside the ESL teaching community could judge a writer's lack of grammatical incompetence:

Working on a test for a [PROFESSIONAL] certification, [...] we sent the materials out for comment from the people who really see them, [AND] they were scathing in their responses, even though they could understand the message about errors. So I, I am, from having, had that experience, this is why I am aware that as a marker I am a different reader from the people in the world, because we had to make our standards according to how "If these people were going to be certified, how will they be perceived in the real world as practitioners." So that's where that comes from. I have seen how harsh people are.

7.5 Summary

In conclusion, the processes Jo followed in arriving at the key criteria in his scoring scheme illustrate how he combined his assumptions about the purpose of a test, and his assumptions about tests takers, with ideas of language development. This is particularly clear in how he established "sophistication" as a criterion. However, while Jo established the broad guidelines for his scoring scheme by relying on his ideas of language development, he generated more specific rating criteria with reference to the corpus he was rating, through repeated compari-
sons of essays, rather than through the application of pre-existing rating schemes (as in the cases of Chris and Sam) or through the application of norms external to the corpus itself (as in the case of Alex). Such a procedure of establishing scoring criteria allowed Jo to be flexible without sacrificing consistency, since he did not complete his scoring session until he was satisfied that compositions to which he gave identical scores were comparable based on the textual qualities that he had used in his scoring scheme.
Chapter 8: Discussion

In exploring the processes raters of ESL compositions in my study followed in constructing scoring criteria and in relating these processes to aspects of their personal and professional backgrounds, I first constructed the four case studies presented in Chapters 4 to 7. The purpose of the discussion in this chapter is to summarize and compare my findings for each rater in the context of the four specific research questions generated in Chapter 3, to evaluate the significance of the findings, and to suggest, briefly, avenues for further research on this issue.

Overall, what I saw emerging most clearly from my data was that in the absence of a scoring rubric (a condition imposed on raters in the Cumming et al. (1998) study precisely to find out how participants constructed scoring criteria) the establishment of a link between performance and proficiency was central to the way raters generated scoring criteria. Each of the four participants provided that link by constructing a developmental trajectory for, presumably adult learners of a second language, on which they could place the writers of the compositions. The data I collected, particularly during the interviews, also revealed that raters constructed their trajectories with reference to their experiences in teaching and (in the case of NNS raters) in learning ESL. Further, in the process of generating specific scoring criteria they combined their ideas of a developmental trajectory with inferences they made regarding the purpose of a test (in this case the TOEFL), and the characteristics of the candidates writing that test. At the same time, while the most experienced raters (Alex and Jo) also relied on their work with test and rating scale development, as well as on their familiarity with theories of language acquisition, their assessment experience per se was helpful principally in the establishment of strategies for reading both individual compositions and a corpus of compositions. With the exception of Alex (who cited the influence of specific rating scales on the way he looked at, though not necessarily judged, compositions), assessment experience exerted limited influence on how participants in my study generated their scoring criteria.
8.1 Research Question A: In what principled way did the raters of ESL compositions participating in the study relate the performance embodied in a written text to a writer's level of proficiency, and how did background factors influence this process?

In the words of Williamson (1993, p. 2), "the context for validation research on holistic scoring begins with an understanding of the distinction between written texts and writing ability," a distinction he traced to Kant. In a similar vein, Bachman (1990, p. 308) stated that "the distinction between language ability and the performance of that ability has been at the same time a central axiom and a dilemma for language testing." What united the participants in the present study was that in the absence of a scoring rubric for their rating task they all attempted to bridge the gap between proficiency and performance by looking for indicators that could place writers on a developmental trajectory, and to assign a score to a composition based, essentially, on its writer's position on that trajectory.

Yet, while participants in the study followed identical procedures, differences emerged between them through differences in the way they defined both language proficiency and the process of acquiring language proficiency. Such differences, in turn, carried over into the way the participants viewed, and compensated for, the impact of situational factors on performance (such as topic effects or time pressure).

Alone among the participants, Sam took the view that language proficiency reflected a person's intellectual development in general. Looking at the influence of background factors, Sam's attitudes were determined principally by his learning and teaching experiences prior to his coming to North America. Since he felt that it was the curriculum followed in his native country that had enabled him and his peers to successfully acquire English, he was confident in relying on the educational concepts that had produced that curriculum. These concepts centred on the notions that a) language ability was a measure of intellectual ability in general, something which Sam confirmed through his personal experiences learning English and writing language tests, and b) that intellectual development could be inferred from a language performance. As a consequence of his far-reaching definition of proficiency, Sam not only used a wide range of criteria to identify a writer's position on a developmental trajectory, but also
found it possible to compensate for situational factors that may have negatively affected performance. In addition, Sam relied on the existence of a developmental trajectory not only to estimate a writer’s level of proficiency; but also to infer her potential.

Chris, being a native speaker of English as well as an experienced teacher of ESL, took his idea that writers could be placed on a developmental trajectory from two sources: a) his observations of how writers whose proficiency he could ascertain through continuous classroom observation performed on tests and b) placement testing, which forced him to think in terms of a developmental trajectory for writers who had to be fitted into a curriculum for writing instruction. Additional ideas on the development of language ability came from Chris’ understanding of communicative competence, and, since that concept does not associate language ability with intelligence, Chris, unlike Sam, interpreted proficiency as communicative competence, a view at once more restricted and more holistic. Consequently, he refused to take situational factors into account in the process of rating compositions, regarding them only as complicating factors in an already complex process. In addition, Chris refused to assess a writer’s future potential on the grounds that this would reduce the reliability of his assessments, restricting his assessment to inferring a writer’s current level of development from her performance.

Alex, meanwhile, possessed extensive experience with standardized assessment, yet he developed his attitudes towards the meaning of writing performance principally from his teaching experiences. Not only had he taught both high-school and university-level classes (in his native country), which afforded him first-hand experience of how students progressed, but he could, like Chris, observe how students’ performances in classroom tests related to their levels of proficiency. Further, although aware of its critics, Alex referred to the Teachability Hypothesis espoused by Pienemann (1986; see also, Pienemann, Johnston & Brindley, 1988) which held that syntactic and morphological structures in English were acquired in a predetermined order, and the order in which such features were acquired could constitute a developmental trajectory. Like Chris, Alex did not associate language ability with
intellectual ability in general, but linked it to the concept of communicative competence. Nor did he take account of situational factors or a writer’s potential in assessing performance, regarding these, like Chris, as sources of unreliability.

Finally, Jo’s views on what a performance represents arose out of his experience with test development, experience which had included piloting writing prompts and establishing scoring rubrics based on the range of writers’ responses to them, and which had given him an idea of what kind of performance writers at various levels of development were capable of. Although he, like Chris and Alex, did not connect language proficiency with intellectual development, his experience with test development and standardized assessment induced him to consider the effects of topics as well as time pressure on performance, simply because he had seen all too often how these factors could impact on performance.

8.2 Research Question B: How did raters of ESL compositions interpret the role of writing prompts in performance-based assessment, and how did background factors influence this process?

In defining her model of the rating process Kroll (1998, p. 223, citing Hamp-Lyons and Kroll, 1997, p. 21) defined a writer as a "complex of experience, knowledge, ideas and emotions … [who] must create a fit between their world and the world of the essay test topic." My data suggest that the participants indeed viewed writers as complex individuals, but only to the extent that the latter succeed in projecting their individuality through their command of both language and ideas, and they made few allowances for the gulf separating a writer’s world from the world of the essay topic. Such a consensus among raters of widely varying backgrounds can, in the present study, be attributed to the fact that each participant regarded the writers of the TOEFL essays as potential university students, and this expectation outweighed their interest in writers’ personal backgrounds. Instead of viewing topics as potentially discriminating, they generally regarded them as posing challenges which writers aiming to be university students had to be able to overcome. This attitude appeared reasonable to Sam because he expected university students to possess a certain level of intelligence,
to Jo because test development had shown him that candidates at higher levels were distinguished precisely by their ability to master unfamiliar situations, and to Alex and Chris because assessment experience had shown them that good candidates could either modify essay topics or adopt roles or stances to suit their purposes.

What the results of the study seem to indicate in this instance is that differences in raters' backgrounds have little impact on how they view writers, if they agree broadly on the purpose of a test and, consequently, form similar expectations of the candidates writing it. A striving for "objectivity," a shared value among most raters of language tests which arises from subscribing (implicitly or explicitly, Williamson, 1993, pp. 5-6) to psychometric theory, also had a leveling effect in this regard. And while "objectivity" clearly didn't prevent participants from estimating a writer's position on a developmental trajectory from a single performance, it did inhibit them from taking writers' ethnic and socioeconomic backgrounds into account in rating compositions.

A second issue which emerged subsequent to analyzing the verbal protocols was why some raters regarded the essay topics as prompts to elicit a performance, and others as tasks which had to be fulfilled or, even, as questions that demanded an answer. In Sam's case it was his own practice of subverting politically motivated questions as a student that seem to have made him tolerant of writers who drifted away from their topic — as long as their language remained on target. Jo, in turn, relied on his work as a test developer, which had shown him how little consensus existed between writers (and raters) on how to interpret a topic (See Connor & Carrell, 1993 for similar findings). It is no coincidence that Sam and Jo, who regarded essay topics as prompts, were more likely to forgive writers who strayed from the topic. By contrast, Alex and Chris consistently expected writers to fulfil the requirements set by a task; Alex because as a rater he had found the ability to fulfil tasks a good discriminator among relatively proficient students, Chris because he had seen too many perfect off-topic essays that were clearly memorized in the hope of getting a high grade.

All in all, the impact of differences in background on raters' attitudes towards writing
prompts was the most difficult to discern. Although Sam and Jo had an academic background in the humanities while Alex and Chris shared a background in the sciences, none of them felt that this had influenced their attitudes. It appears, instead, that Sam and Jo had developed a greater sense of empathy because of the nature of their encounters with language tests. Sam, being a learner of English himself, had experienced the frustrations of being unable to express himself with the desired precision. Jo, on the other hand, had seen the struggles of writers in the course of his work on test development, where the aim was not to judge the production of the writers but to judge the appropriateness of a particular prompt in light of what writers had been able to do with it. Chris and Alex, by contrast, had experience not with writing or developing tests but with scoring them in contexts where they could take the validity of writing prompts for granted given the amount of effort invested in their development by the institutions they worked for.

8.3 Research Question C: What reading strategies did raters of ESL compositions establish (to deal with both individual compositions and a corpus of compositions) and how did background factors influence these?

Although participants’ scoring procedures did not determine their rating criteria, I had found it necessary to examine the steps they followed sequentially in order to analyze the criteria which they had considered in scoring compositions. The four raters adopted one of two distinct strategies; Alex and Jo took multiple (up to four) passes through the compositions, and sorted them into preliminary piles before deciding on a score, while Sam and Chris followed a read-through approach, reading each composition only once and rendering a score immediately afterwards.

Not surprisingly, this is an area where assessment experience played a key role in shaping the attitudes and behaviour of the participants. Sam used the procedures he had developed in classroom assessment, in response to both time pressure and a high level of confidence in his judgements, which determined that he needed to read each composition only once before giving a score. Alex and Jo followed the elaborate procedures they had found useful in dealing
with large corpora of compositions in the context of standardized assessment. Only Chris
departed from the procedures he normally followed in dealing with placement tests (the
closest that his assessment experience came to the anonymous setting of large-scale, standar-
dized testing) by not sorting and rereading compositions; he attributed this to the lack of
practical consequences that his grades carried in what was a research context, rather than an
authentic assessment situation.

In conclusion, the patterns showed by raters in the present study suggest that rating stra-
tegies, particularly macro-strategies employed in rating a corpus, may represent an aspect of
performance-based assessment which is largely impervious to teaching or learning experience
and must be developed through focused training and practical assessment experience. Looked
at another way, of course, it may be the one area where focused training (in the context of
what Pula and Huot, 1993, p. 256 called an "immediate discourse community"), as well as
explicit instructions at the outset of a rating session, can successfully modify raters’ beha-
viour to suit the purposes of a specific assessment context.

8.4 Research Question D: By what process did raters of ESL compositions construct specific
rating criteria, and how did background factors influence this process?

As discussed under Research Question A, in the absence of a scoring rubric, the initial step in
the participants' establishment of scoring criteria involved the assumption that a learning
curve could be established for ESL writers, from which specific criteria could be generated to
show the position of a writer on a developmental trajectory. The precise nature of that curve
depended, first, on assumptions about language proficiency and language learning. To this had
to be added assumptions about test takers and test use, issues that were discussed under
Research Question B, which dealt with the role of writing prompts in performance-based
writing assessment. Raters also had to make decisions about where the end-points of their
learning curves were, particularly about their perceptions of the requirements for the highest
score. In addition, they had to take into account the instruction that they were to score
papers on a 6-point scale, since this helped them to parcel out the learning curve into 6
roughly equal parts.

All of the participants in the study followed these steps, all agreed that the purpose of the TOEFL essay was to determine whether a candidate for admission to a North American university had the requisite proficiency in English, and all followed the same instructions given at the outset of the rating session, namely to assign scores to compositions on a 6-point scale, and to do so without reference to any existing rating scale. That differences in their scoring criteria arose in spite of these parallels can be attributed to the following factors: participants differed in their perceptions of language proficiency, they differed in their assumptions of how language could be acquired, and they differed on how the end-points of a learning curve (particularly the upper end) could be defined. The first two of these sources of variability have already been discussed in relation to Research Question A. As for the third, all participants in the study expected candidates to provide a clear and well argued answer that fully satisfied all the requirements of the prompt, but while Sam and Alex did not expect native-like control of formal aspects even from candidates at the highest level, Jo and Chris did, and gave 6s only to essays that were, essentially, error-free.

8.5 Implications

The finding that raters in the study constructed rating criteria by combining assumptions about language proficiency and language learning with information about test use, test taker characteristics, specific instructions given at the outset of the rating session, and, in some cases, the qualities of the corpus they were rating, has several implications. First of all, given the nature of the information that raters accommodated, the process of establishing rating criteria must be repeated anew for every corpus of compositions that raters assess. Furthermore, while raters in the present study were explicitly instructed not to rely on a scoring rubric, the two most experienced participants did suggest in their interviews that even if they had been instructed to use a scoring rubric they would have had to specify their scoring criteria through taking additional factors into account. This suggests that rating scales are not
the sole determinants of writing quality in raters' judgements, but are regarded by raters as only one (even if possibly the weightiest) of several factors that they must take into consideration. Such a hypothesis could explain why it is that even with a specific scoring rubric inter-rater reliability can usually be achieved only through extensive discussion of rating criteria and the establishment of anchor papers within a holistic scoring group.

The issue that needs to be addressed next is to what extent variability in judgements can be ascribed to variability in raters' backgrounds. Although the data in the present study were generated through reliance on a very small number of case studies, and are hardly amenable to generalization, they can be used to highlight a range of factors relevant to variability, whose relative importance could subsequently be explored through a larger sample of ESL composition raters. Even my limited data have shown, for example, that the factors that I had isolated in the initial literature review need to be reassessed. Differences among raters based on ethnicity, culture and mother tongue are certainly important, and these background factors may be viewed as extreme cases of "extended discourse communities" (Pula & Huot, 1993). Academic background, however, did not emerge as a significant factor, and assessment experience was likewise limited in its impact, influencing mostly rating strategies, but not the establishment of scoring criteria. Nor is there any support in my data for the idea advanced by Pula and Huot (1993, p. 252) that raters compare the compositions they are rating to a model that they had internalized through their reading experiences.

The most significant sources of variability in the case of the four participants in the present study lay, clearly, in teaching experiences and, in the case of the NNS raters, learning experiences. Such experiences were at the root of raters' conceptions of a developmental trajectory for language learners, which represented the initial step in their construction of rating criteria. Although theoretical positions on language proficiency (such as Communicative Competence) and language learning (such as the Teachability Hypothesis) could confirm raters' stances on these issues, such stances in the case of the present study were rooted more fundamentally in personal observations in the language classroom. Such a finding would
lend strong support to recent efforts to combine insights into language testing with insights into second language acquisition (best exemplified by the papers in Bachman & Cohen, 1998), in view of the demonstrable failure of most existing rating scales to properly take account of ideas of language development (Brindley, 1998), and the corresponding difficulty of using existing theoretical models of language development (e.g., Pienemann, et al., 1988) to design viable language tests. In addition, the fact that the raters in the study were most heavily influenced by their teaching (and, if applicable, learning) experiences may additionally explain why training in assessment procedures, indeed assessment experience itself, does not in itself ensure that a group of raters can render consistent judgements using the same rating scale. It may be better, instead, to quiz potential raters on their attitudes to the acquisition of language proficiency, indeed to the nature of language proficiency itself, in order to ensure not only consistency in judgements but also that the same construct underlies these judgements.

In view of the findings, and of recent trends in language testing which have just been referred to, the present study could be usefully followed up with a study of how raters assess ESL compositions when they are instructed to use a scoring rubric. Will they apply the rubric to the exclusion of other considerations, or will they confirm the hypothesis I advance above that raters regard the rubric as yet another factor, even if a highly significant one, which has to be weighed along with their ideas of a learning curve for ESL writers in establishing specific scoring criteria? It would also be useful to get raters to rate writing prompts using different scoring rubrics to see if they react differently to them. If so, what are the characteristics which can account for differences: are they rooted in the scoring rubrics themselves (e.g., rubrics focusing on performance outcomes, such as the ACTFL scoring guidelines or the IELTS proficiency bands, versus rubrics that are based on theoretical models of language development)? Until such questions are answered (and they couldn’t be in the present study due to the special circumstances under which the raters operated), even if reliability in judgements can be secured through specific instructions and focused training, the central question of what those ratings represent will not be answered.
References


Chalhoub-Deville, M. (1995b). Performance assessment and the components of the oral construct across different tests and rater groups. In M. Milanovic & N. Saville (Eds.). *Performance testing, cognition and assessment. Selected papers from the 15th Language*


Appendix A: Letter of consent signed by participants in the study

Dear ________________  January 8, 1999

I am writing to ask you whether you would be willing to participate in a research study entitled, "The influence of background factors on decision-making behaviour in the performance-based assessment of ESL writing ability." The purpose of my research, undertaken in partial fulfillment of the requirements for the Master of Arts degree at OISE/UT, and supervised by Dr Alister Cumming, is to relate the backgrounds of four experienced raters to their behaviour in marking a sample of ESL compositions. This will complement the ETS study, "An investigation into raters' decision-making and development of a preliminary analytic framework for scoring TOEFL essays and TOEFL 2000 prototype writing tasks," in which you participated during June-September, 1998. I have permission from ETS to carry out the study and will report my findings to them.

In setting up my study I have decided to choose four experienced raters of ESL compositions who - as a group - would exhibit maximum variability in terms of their L1, academic background, teaching experience and experience with both large-scale writing assessment and specific rating scales. The reason I would like to include you in my study is that your participation would enable me to achieve the ideal balance in these variables.

If you are willing to participate in the study, I will ask you to allow me to use the data gathered from questionnaires and think-aloud protocols during the ETS study, as well as to amplify this information in two interviews which will elicit the following additional information:

- your personal background and professional experience
- your interpretive comments on a sample of your think-aloud protocols produced during the ETS project in June-September, 1998

- your comments on my analysis of your behaviour

The first two items will be elicited in the course of a single, taped interview which I estimate will take three hours to complete. The timing of the interview will be arranged according to your convenience and while I will not be able to pay you an honorarium, I will be happy to cover your travel costs (or, alternatively, arrange for the interview to take place at a location other than Toronto, should that prove to be more convenient to you). The follow-up interview will involve your reading a draft of my thesis and verifying it for accuracy and for the fit between my interpretation and your own.

Your participation in this study will be voluntary and you will be free to withdraw from the study at any time, as well as free to withdraw permission to use data from your answers to questionnaires and from your think-aloud protocols collected during the original ETS study. Your involvement in the study will, moreover, be kept confidential, and your identity will be concealed by a pseudonym (different from the one that you had used for the ETS project). Further, the aim of my study is not to evaluate your performance in any way, but, rather, to understand it more fully, so I will avoid any statements in my analyses that are judgemental of your professional competence.

At the conclusion of my analysis I will provide you with a detailed profile which will attempt to relate your personal and professional background to your ratings of a sample of ESL compositions.
If you agree to participate, please sign the statement on the following page and return it to me. Should you require further information please contact me by telephone (H: (905)-271-7134; O: (416)-978-6838) or email (my address is: uerdosy@oise.utoronto.ca).

Sincerely

M. Usman Erdosy
Reply form

I have read Usman Erdosy's letter of _______ describing the aims of his research project, "The influence of background factors on decision-making behaviour in the performance-based assessment of ESL writing ability," and believe that I understand its purpose, contexts and means protecting my personal identity, privacy and reputation. Consequently, I agree to provide the data along the lines described therein, on the condition that I may withdraw from the study, as well as withdraw permission to use data already collected in the course of the study, at any time.

(signature)

Name: __________________________

Date: __________________________
Appendix B: Transcription conventions

I have used the conventions developed for the Cumming et al. (1998, pp. 80-81) study for the transcriptions of the think-aloud protocols and the interviews:

**Symbols**

( ) uncertain transcription  
`x` incomprehensible item, 1 word only  
`xx` incomprehensible item of phrase length  
`xxx` incomprehensible item beyond phrase length  
`...` pause of 5 seconds or more  
`-` incomplete word  
`—` interrupted or incomplete sentence  
" " text read directly from the original composition

For standard hesitation markers or exclamations, I have used:

- **Yeah** for colloquial yes, yeh, ye, or ya  
- **O.K.** for okay  
- **Uh** for short hesitation sounds  
- **mmm** for long hesitation sounds  
- **oh** for exclamations

**Symbols added for the present study**

- **[SMALL CAPS]** To paraphrase statements made by participants in the interviews in order to shorten quotations  
- **[QUOTE]** To indicate direct quotes from either essay prompts or the TOEFL essay themselves  
- **[...]** To indicate ellipsis (on my part, to shorten direct quotations).

I did not edit participants’ transcripts for grammar in order to avoid imposing any interpretations on the raw data. For a sample coded transcript, see Appendix C, below.
Appendix C: Sample coded protocol

(NB: Codes appear in bold capitals above the utterances; the symbol "/" is used to indicate the boundaries of an utterance corresponding to a code while the symbol "||" is used to denote the boundaries of segments within the protocol)

\[ P / IL1 / JL1 \]
Essay number 65. It's typed. It appears to be in four extremely short paragraphs, but at least

\[ JR7 / IR2 \]
an attempt at uh, making, or organizing it into paragraphs. Just looking at the topic sentence.

\[ IS2 || JR1 || IS2 || JL1 \]
[QUOTE] O.K. so nice and clear at the beginning. [QUOTE] Good. Although short, but at least

\[ JR1 || IS2 || JR4 \]
an attempt at an introduction [QUOTE] O.K. they have used a time, or not time, importance

\[ IS2 || IL2 || IS2 || IL3 \]
marker. [QUOTE] is misspelled. [QUOTE] I think they are trying to say [...] looks like a typing

\[ IS2 || JR1 \]
mistake [QUOTE] So they've made the point and at least come up with an example to support

\[ IR2 || IS2 || IL2 || IS2 \]
it. The next body paragraph, [QUOTE] There is an article mistake there. [QUOTE] O.K. so

\[ JR4 || IS2 \]
there it is, sort of, a bit incoherent, they are talking about [QUOTE] O.K., although at a glance

\[ JR4 \]
appeared to be in fairly coherent paragraphs, in, in fact it's not, at least, I mean they have

\[ JR1 \]
used [TRANSITION EXPRESSIONS] There seems to be some awareness of providing examples

to support the points, although the examples are weak and not, uh, not developed at all. …

\[ JR4 || JS7 \]
However, I think this one would get a 2. … Actually I am going to change my mind on that

\[ JR1 \]
because they haven't really, I'm just looking it over again and I'm thinking that they, although

\[ JR3 || JS7 || \]
they have listed three points, they haven't related it to the topic, uh, precisely, so, a 1.