ETS®
TOEFL®

Research Reports

Report 70
January 2004

Exploring Variability in Judging Writing Ability in a Second Language:

A Study of Four Experienced Raters of ESL Compositions

M. Usman Erdosy

Exploring Variability in Judging Writing Ability in a Second Language: A Study of Four Experienced Raters of ESL Compositions

M. Usman Erdosy

Modern Language Centre, Department of Curriculum, Teaching, and Learning
Ontario Institute for Studies in Education

RR-03-17

Abstract

Variability in judgments of ESL compositions is inherent in the view that raters are "readers" with prior experiences. Such a view, however, obliges researchers to understand how personal background and professional experience influence both scoring procedures and scoring criteria. These issues were explored by asking four raters to construct scoring criteria while assessing corpora of 60 TOEFL essays without the aid of a scoring rubric, and to discuss their procedures and criteria in follow-up interviews. The study revealed key points in the decision-making process, where raters' behavior diverged, and examined the impact of prior experience on these. The identification of such divergences, and potential explanations for them, were undertaken to lay the foundations for a principled explanation of rater variability.

Key words: ESL writing, ESL writing assessment, language assessment, rater variability, second language writing, writing assessment

The Test of English as a Foreign Language^{TM} (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service® (ETS®) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations. GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Committee of Examiners. Its 13 members include representatives of the TOEFL Board, and distinguished English as a second language specialists from the academic community. The Committee meets twice yearly to oversee the review and approval of proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Committee of Examiners serve three-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Because the studies are specific to the TOEFL test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language and applied linguistics. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

To obtain more information about TOEFL programs and services, use one of the following:

Email: toefl@ets.org

Web site: www.toefl@.org

Acknowledgments

This monograph represents a revised version of my master's thesis (Erdosy, 2000), written at the Modern Language Centre, Department of Curriculum, Teaching, and Learning, Ontario Institute for Studies in Education. I am very grateful to Professor Alister Cumming, my thesis supervisor, and to Professor Merrill Swain, the second member of my thesis committee, for their comments both prior to and subsequent to the submission of my thesis in partial fulfillment of the requirements for the M.A. degree in second language education. I am equally grateful to the participants in the study for providing much valuable data, to Educational Testing Service (ETS) for allowing me to use data from a research project concerned with the behavior of raters of ESL compositions (Cumming, Kantor, & Powers, 2001), and to the anonymous reviewers who made valuable suggestions for improving the manuscript in both clarity and focus.

Table of Contents

List of Tables

Introduction and Research Questions

Although performance-based assessment has become the method of choice in judging writing ability in English as a Second Language (Kroll, 1998, p. 221), variability in the scores raters assign to ESL compositions remains a concern for validation studies. Such variability has been demonstrated between raters varying in their level of experience (Cumming, 1990; Weigle, 1994), teaching focus (Santos, 1988), or mother tongue (Brown, 1995; Chalboub-Deville, 1995a, 1995b), as well as in the judgments of individual raters over time (White, 1984). It may be minimized through refinements in rating scale descriptors and writing prompts (Alderson, 1991), and through training raters in the use of specific rating scales (Weigle, 1994). It may even be viewed as natural if raters are regarded as "readers" (as in Huot, 1990; Janopoulos, 1993; Kroll, 1998; Purves, 1984). However, as Milanovic, Saville, and Shuhong (1995, p. 93) have argued, such variability reflects an underlying disagreement over the ability raters aim to measure, and is thus inextricably linked to the construct validity of performance-based writing assessment.

Faced with rater variability, the advocates of performance-based writing assessment have approached construct validation from several angles. There have been studies linking variability in raters' judgments to variability in specific textual characteristics in candidates' responses (see Cumming, 1997, for a summary of results). Scoring rubrics have been examined to find out if raters using them differed in their levels of severity or consistency, or if all the points on a rating scale had an equal chance of being used (e.g., McNamara, 1996; Tyndall & Mann-Kenyon, 1996). There have even been attempts to incorporate construct validation into the process of rating scale construction, through empirical observation (Fulcher, 1997; Upshur & Turner, 1995) and through rationales grounded in theories of language acquisition (Bachman & Cohen, 1998) or models of communicative competence (McNamara, 1996). However, such studies can only be viewed as useful starting points in the process of explaining why raters behave the way they do. If raters of compositions are indeed "readers," relying not on absolute standards, but on their experiences, expectations, and purposes, to evaluate writing ability, the source of such experiences, expectations, and purposes will play a key role in explaining rater variability. Consequently, it becomes important to understand how raters' prior knowledge and expectations are rooted in personal background and professional experience, and how they translate into specific scoring criteria during the rating process.

It is in this context, and in view of the failure of studies that took a one-dimensional view of both the rating process and the influence of background factors, that I conducted my research, in three stages. First I analyzed data previously generated through questionnaires and concurrent verbal protocols by a project investigating raters' decision-making behaviors (Cumming, Kantor, & Powers, 2001), in order to provide detailed descriptions of the way four experienced raters rated corpora of 60 TOEFL essays. Subsequently, through interviews I designed based on my analysis of the protocol and questionnaire data, I explored the influence of background factors on participants' rating processes. Finally, I drew contrasts in the way individual raters constructed scoring criteria while performing a specific rating task, and explored the influence of personal and professional backgrounds on the processes they followed, considering these to be the initial steps toward a principled explanation of variability in rater judgments.

It is the key findings that I will present here, starting with my research questions, followed by a review of related research concerning the influence of background factors on the decision-making behaviors of raters of ESL compositions. Although these have failed to yield consistent findings that could satisfactorily explain how raters of ESL compositions actually constructed scoring criteria, the studies have tentatively identified formative influences on raters' strategies and judgments and were useful in selecting participants and drawing up questions probing them about aspects of their backgrounds. Subsequently, I will describe my research design, including choice of participants, research instruments, and methods of analysis. Then, I will present the steps participants in the study followed in constructing scoring procedures and scoring criteria and the influence of background factors on this process, based on my analyses of the participants' think-aloud protocols, questionnaires, and interviews and structured according to the four research questions already posed. Finally, I will discuss the implications of the study's findings.

Research Questions

Aiming to complement the 2001 study of Cumming et al. concerning the decision-making behaviors 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 did raters of ESL compositions follow in constructing scoring criteria to assess a corpus of compositions in an experimental situation where they had no predetermined criteria to rely on, and how can these processes be related to their personal and professional backgrounds?*

To sharpen my focus, I then followed Kroll (1998, p. 223) in identifying “reader (= rater),” “written text,” “writer,” “writing prompt,” and “scoring procedure” as the critical constituents in the process of direct writing assessment. Using this model, I generated four research questions — exploring the attitudes of the participants in my study to, in turn, the written text, the writer, the writing prompt, and the scoring procedure — which, I felt, reflected 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 after analysis of the concurrent think-aloud protocols generated in the first part of the study. To begin with, I followed Kroll (1998, p. 223) in breaking down “scoring procedure” into two distinct aspects: scoring criteria and reading strategies. Although reading strategies probably had no direct bearing on the construction of scoring criteria, participants’ rating sessions in the present study showed distinct stages resulting from their approach to reading both individual compositions and the entire corpus of compositions, and it was impossible to describe their behavior without the contexts provided by these stages in their procedures. Consequently, I decided to address this issue in the first research question:

> *What reading strategies did the raters of ESL compositions participating in the study establish to deal with both individual compositions and a corpus of compositions, and how did background factors influence these?*

Next, I found that my participants’ attitudes toward a written text represented answers to a more fundamental question, concerning the way one can relate “performance” (in this case embodied in a single text produced under carefully specified constraints) to “proficiency.” Addressing this issue constituted my participants’ initial step in their processes of constructing scoring criteria; consequently, I phrased my second specific enquiry as follows:

> *In what principled way did the raters 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?*

Subsequently, examining the perceptions of my participants concerning the writers of the compositions they were rating, I found that these 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 the performance of the writers, and I could subsume this aspect of their behavior 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, eventually articulated in the following manner:

> *How did the raters participating in the study interpret the role of writing prompts in performance-based assessment, and how did background factors influence this process?*

Finally, in my preliminary analysis of protocol data I saw scoring criteria emerging not only from assumptions concerning the meaning of performance but also from specific information that raters gathered while taking part in the rating session. Such considerations produced the final research question, which treated the construction of scoring criteria as a complex process in which raters' personal and professional backgrounds played a major role:

> *What specific scoring criteria did the raters participating in the study generate in the absence of a scoring rubric, what information did they heed during this process, and how did background factors influence their choices?*

Related Research

Each of my research questions concerned both the process raters of ESL compositions followed in constructing scoring criteria, and the influence of background factors on the steps involved in that process. In exploring the former, I found existing models of the rating process (particularly Kroll, 1998) as well as existing categorizations of rater behavior (particularly Cumming, 1990) useful in constructing my hypotheses. By contrast, explanations relating background factors to the behavior of raters of ESL compositions were unsatisfactory because they reduced not only background factors but also the complex process of constructing, or applying, scoring criteria to single dimensions. Nevertheless, several background factors have emerged from the existing literature as potentially significant. Consequently, I decided to set the context for addressing the second part of my four research questions by critically examining current hypotheses concerning the interaction of background factors in shaping the attitudes and behaviors of raters of ESL compositions. Such a literature review could also assist me in selecting participants for my study, and in probing potentially significant aspects of their backgrounds through interviews. 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 (NSs) of English, of oral proficiency, as well as of proficiency in other languages, if these provided relevant insights.

Cultural Background

Three factors — cultural background, mother tongue, and professional experience — appeared in studies to exert particularly strong influences on rating behavior, although none could explain it in isolation. Among these, the importance of cultural background was perhaps the first to be appreciated, thanks to Kaplan's (1966) assertion that writers' rhetorical patterns varied across cultures. It was following his seminal work, originally intended to help improve learners' control of English rhetorical devices, that contrastive rhetoric was established as a distinct field of study. The cultural divisions established by Kaplan (e.g., "Oriental," "Romance") appear, in retrospect, to be too broad, and the approach is also dangerously open to stereotyping (Raimes, 1998). However, certain rhetorical conventions may well be culturally determined. For example, a survey of holistic writing assessment of ESL compositions (Tedick & Mathison, 1995) yielded a generally positive correlation between the score awarded and raters' judgments about the effectiveness of framing, measured by the degree to which the raters could successfully predict the topic and the pattern of development of an essay after reading only the first paragraph.

Similarly, a study of peer reviews in scholarly journals (Belcher, 1995) found that a clear thesis statement at the outset was considered essential by readers. From such findings it appears that a linear form of reasoning could usefully be associated with writing in English, just as Kaplan had suggested.

Judgments related to content and tone can be similarly affected by cultural considerations. Witness the preference among educated NSs of English for detachment, particularly in academic writing, at the expense of personalizing issues (cf. Basham & Kwachka, 1991) or the preference for using an indirect strategy to present bad news in both academic (Belcher, 1995) and business (Guffey & Nagle, 1997) transactions. Clearly, then, there is a compelling need to relate rater judgments and strategies to cultural background. Yet it should be stressed that no single background factor, not even cultural background, can alone account for rater behavior. Note, also, that attitudes to rhetorical conventions, however deeply they are 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 than adherence to a particular rhetorical pattern by all raters in their study, regardless of their cultural backgrounds. Tedick and Mathison (1995) likewise found that even the failure to frame a composition could be overlooked in certain circumstances. Thus, while cultural background is potentially significant, it is only one of a range of factors affecting rater behavior.

Linguistic Background

The influence of raters' linguistic background emerges most clearly in the contrasting attitudes of NS and nonnative-speaker (NNS) raters of ESL compositions, and given the internationalization of English and the increasing number of NNS assessors, this difference 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 than NSs, and explained their findings by referring to the considerable time and energy NNS assessors had invested in learning a language, which led them to attribute errors to a lack of commitment on the learners' part (Santos, 1988, p. 85). Others have found otherwise; for example, Brown's (1995) study of the assessment of the proficiency level of Japanese-speaking tour guides revealed that NNSs were more lenient in every respect than NSs, except in judging the mastery of politeness strategies.

In trying to account for such inconsistencies, one may again criticize a preoccupation with a limited set of factors, at the expense of others that may have also influenced behavior. For example, the NS judges in Santos's (1988) study were college professors and their ratings had no practical consequences for the writers because they were used only for research purposes. By contrast, in Brown's (1995) 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 not only for applicants but also for the reputation of their profession. Under such circumstances it is not surprising that NNS assessors in Santos's study turned out to be lenient, and in Brown's study they 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 NS and NNS reactions to the single dimension of rater severity. These are two key lessons to be learned from studies conducted by Chalhoub-Deville (1995a, 1995b) concerning varying perceptions among Arabic speakers of proficiency in Modern Standard Arabic (MSA), the lingua franca that enables the speakers of various Arabic dialects to communicate with each other. Although in this situation concepts of "NS" and "NNS" do not strictly apply, Chalhoub-Deville found that Arabic speakers living in Lebanon equated learner proficiency in MSA almost exclusively with linguistic accuracy, whereas those living in the United States valued overall communicative quality as much as accuracy. The reason for the difference, according to Chalhoub-Deville, 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 these studies is that raters' linguistic status is important, but cannot be reduced to a simple NS-NNS contrast without examining the social role that a language such as English may play, particularly in situations where it is not the first language of the majority of speakers. Moreover, linguistic status 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 raters' behavior.

Professional Experience

Professional experience emerges as an important variable from studies in two ways. To begin with, raters may be exposed to very different learner populations as teachers, leading them to form different expectations from learners. This is most easily demonstrated through documented differences between the rating behaviors of teachers of English and the rating behaviors of teachers of English as a Second Language. Teachers of English 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. Such differences in judgment may be attributed to cumulative context effects. Research (e.g., Miller & Crocker, 1990; 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 likely to react more negatively to NNS errors than would ESL teachers accustomed to ESL learners' errors. Nor should we be surprised that teachers of English to NSs may overlook positive discourse features in otherwise weaker essays because 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.

Previous teaching experience may also become an important factor if raters are prone to carry certain teaching concerns into the rating process. I was particularly interested in observing if participants in the present study with teaching experience were focusing their evaluations on textual features that play an important part in their teaching curricula, on the assumption that raters with teaching experience are accustomed to giving feedback as a means of encouraging learners to learn (Prior, 1995).

A second key dimension of variability is level of assessment experience. There are, not surprisingly, differences between professional and lay raters in general, even if, as elsewhere, the findings from research are inconsistent. For example, Barnwell (1989) finds professional raters to be more tolerant of language errors and suggests, not unreasonably, that their exposure to the widest possible range of linguistic ability has enabled them to put learners into a more realistic, thus more forgiving, perspective. Others (e.g., Galloway, 1977; Hadden, 1991) come to the opposite conclusion. Inconsistencies even extend to individual aspects of language ability being judged: Hadden identifies pronunciation as an area where professional raters' judgments are more severe, whereas Brown (1995) concludes that professional raters in her study were harsher in all areas of proficiency except pronunciation. Although these differences are of little immediate significance to a study concerned with the behavior of experienced raters of ESL compositions, the findings are useful in deciding just whose judgments are to take precedence. What is the value, for example, of the balanced assessment of proficiency normally rendered by experienced raters, if, as in Brown's study, the inability to fulfill a vital task renders a candidate inadequate in the eyes of his peers in spite of the general excellence of her performance? Although such a question is surely embedded in Messick's (1989) expanded concept of construct validity, it remains to be fully addressed.

More usefully for the present study, the influence of training and rating experience on the behavior of the more restricted population of raters of ESL compositions has also been dealt with abundantly in recent studies. Key findings, in short, include 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, focused training can help raters apply scoring schemes consistently (Connor & Carrell, 1993; Weigle, 1994). On the other hand, the benefits of training may be short term (Lumley & McNamara, 1995), and training cannot eliminate context effects (Hughes, Keeling, & Tuck, 1983). In addition, as Hamp-Lyons (1990) 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." (p. 81).

Finally, one reason for divergent ratings may be raters' familiarity with different rating scales in assessing compositions, whether in class or through standardized tests. As Pula and Huot (1993) point out, membership in a holistic scoring task group represents an important element in raters' backgrounds, which may influence the way the raters look at compositions. Consequently, participants for the present study were chosen partly because in the questionnaire concerning their professional backgrounds, completed immediately after generating think-aloud protocols, they indicated experience with using specific rating scales. 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.

Implications

My first conclusion based on the reviewed literature is that neither raters' backgrounds nor raters' judgments are reducible to a single dimension. At the same time, there are few useful models relating a range of background factors to rater behavior. 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 their reading and writing experiences, 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. Overall, Pula and Huot concluded that raters of English compositions internalized a model of good writing, principally based on their reading experiences, and compared the compositions they were rating to such a model.

Although Pula and Huot (1993) provide a useful exercise in model building, and employ a sound methodology in combining coded verbal protocols with confirmatory interviews, their findings appear too simplistic to be applicable to raters of ESL compositions. Internalizing a static model of good writing may be acceptable when raters can assume that the writers of the compositions they are rating have reached the limits of their potential — that is, when the writers are NSs of English — but when the writers are learners of a second language, a more dynamic model is likely to be called for. In any case, the assertion that all the raters in their study internalized the same model — implied by the statement that each rater emphasized "content and organization as their primary criteria in scoring" (p. 252) — is grossly inadequate. 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 whereas 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, I was able, in reviewing the published literature, to identify three potentially significant sources of variation in the backgrounds of raters of ESL compositions: cultural background, linguistic background, and professional experience (itself manifested in at least two ways). How this information helped me to select participants for this study and elicit data from them through interviews is discussed in the research design section, presented next; to what extent my choice of these four background factors helped explain variability in the behavior of the participants in the study is, in turn, addressed in the results section.

Research Design

Choice of Participants

Having identified three important potential sources of variability — cultural background, mother tongue, and professional experience — in raters' backgrounds through a literature review, I chose participants in such a way as to maximize variability along these dimensions. The four eventually chosen were among the 10 who had taken part in the Cumming et al. (2001) study in the dual roles 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. study in their own homes or offices, they met regularly to discuss how the data generated by the project could be interpreted. Because they provided additional data in the present study, by granting an interview exploring the influence of their backgrounds on their rating behavior during phase I of the Cumming et al. study, I asked them to sign additional forms, in which they gave their informed consent to my use of information relating to their backgrounds on condition that their identities be concealed and that they be free to withdraw from the study at any time.

The first participant, Sam, was an NNS of English from Europe, who had come to North America two years before the study was conducted. He had been a certified teacher of ESL in his native country for six years. His experience in assessing ESL compositions was limited to classroom evaluation, his only encounter with standardized assessment having been in the role of a test taker. The second participant, Chris, was an NS of English. Besides teaching ESL at the university level for eight years by the time the study was conducted, his experiences in evaluating ESL writing had included not only classroom and portfolio assessment but placement testing as well. On the other hand, he had no experience with standardized testing. Alex, the third participant, was an NNS from Asia. Besides teaching English and English for Specific Purposes (ESP) at the university level for 12 years in his native country, and conducting classroom evaluation and placement testing, he had worked as a composition rater for a nationwide English examination authority in his native country and had also been involved in rater training. Finally, Jo was an NS who had extensive experience not only in teaching ESL at the university level, but also in teacher training, test development, and the construction of scoring rubrics.

Together, these four participants represented a wide range of variation in terms of the background factors identified as potentially significant: Sam and Alex were NNSs (who differed greatly in the extent to which they had used English in their native countries), Chris and Jo were NS with a North American cultural background; Sam and Jo had a background in the arts, Chris and Alex in the sciences; Alex and Jo had extensive experience with large-scale, standardized assessments, Sam and Chris had none. In addition, all four had mentioned being influenced by different scoring rubrics: Sam by an unlabeled, 5-point holistic scale in his country of origin, Chris by a 6-point rating scale for placement testing, Alex by the scoring rubrics published by Hamp-Lyons (1991) and by Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981), and Jo by the scoring guide for the Test of Written English (Educational Testing Service, 1989). Conversely, the four participants all had in common certification in teaching ESL, extensive experience in teaching ESL writing, and graduate work at the same North American university; in addition, all had participated in the research project on raters' decision-making behavior (Cumming et al., 2001) that my study aimed to complement. One of the participants was male and the other three female; as a further measure of masking their identity I assigned androgynous pseudonyms to them, and referred to them (and raters in general) in the study with masculine pronouns, reserving the use of feminine pronouns for writers whose compositions were assessed.

Data Collection

Data for the study were obtained from four sources. Information on how raters constructed scoring criteria was extracted from concurrent verbal protocols and from the scores participants assigned to compositions, with positive and negative comments concerning textual features associated with scores on a 6-point scale forming the core of my analysis (which, for reasons arising out of the nature of concurrent verbal protocols, to be discussed below, was largely qualitative in nature). The influence of background factors on raters' strategies and judgments was then explored through interviews, a questionnaire, and (scarce) references by participants to their backgrounds in the concurrent verbal protocols. In addition to thus using multiple sources of evidence to facilitate triangulation (Miles & Huberman, 1984, pp. 234-235), 1 incorporated member checks (ibid., p. 242) into my analysis in two ways. First, my analysis of how participants constructed their scoring criteria, based on the protocols, was presented to them in, and used to design the questions for, the interviews. Then, once my analysis of the influence of background factors on constructing scoring criteria was written up, subsequent to the interviews, it was also presented to each participant, who thus had a chance to verify or dispute the accuracy of my analyses at two points during my study.

Needless to say, the circumstances under which participants rendered their judgments were not reflective of typical practice in standardized tests. However, I (with Cumming et al., 2001) felt that the absence of a scoring rubric would make it easier to highlight the processes involved in raters' construction of scoring criteria and to link the variability observed in raters' judgments and behaviors with variability in their backgrounds. The absence of practical consequences to raters' decisions, by contrast, was a limitation imposed by the experimental nature of the study, and its impact on the findings will be assessed in the concluding discussion.

Concurrent verbal protocols. Concurrent verbal protocols were produced by the participants while rating randomly sequenced corpora of 60 TOEFL essays as part of the research project on raters' decision-making behavior (Cumming et al., 2001). The essays had been written in response to four topics during a (then) recent administration of the test and represented all six scale points on the usual scoring rubric for TOEFL essays (as assigned by the original assessors at ETS). They formed a corpus of 144, out of which participants were assigned randomly selected sets of 60. Neither the identity of the authors nor the scores assigned to the compositions during the actual administration of the test were revealed to the participants. Instead, they were asked to assess the compositions independently, and while they were instructed to assign grades using a 6-point scale, they were also instructed not to rely on the usual scoring rubric for TOEFL essays. Following their rating sessions, participants were paired and asked to transcribe each others' verbal protocols.

The principal advantage of concurrent think-aloud protocols in the context of this study was that they provided evidence of cognitive processes that was not colored 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. 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 (e.g., Connor & Carrell, 1993; Cumming, 1990; Huot, 1993; Pula & Huot, 1993; Vaughn, 1991).

This is not to say that the method is without problems. For example, informants may 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., 2001) that yielded the concurrent think-aloud protocols utilized in this study. 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 might have been using a mixed code in my mind. 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. /Alex/[1]

Consequently, following the methods of Cumming et al., I based my analysis of the data on "impressionistic interpretations of patterns and trends," although supplemented by frequency counts of coded behaviors, in a manner to be described below.

One final danger was that concurrent think-aloud protocols could alter the behavior 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, simply to become comfortable with doing concurrent verbal protocols. In such a situation, the triangulation of data from multiple sources was particularly valuable, because erroneous interpretations of behavior based on one source could be identified based on another.

[1] To distinguish excerpts from participants' concurrent verbal protocols and interviews, respectively, the former will be printed in **Arial,** and the latter in **Times New Roman.**

Questionnaires. In addition to providing verbal protocols, all participants in Cumming et al. (2001) had been asked to complete a questionnaire immediately after their rating sessions in order to provide information on their personal and professional backgrounds, as well as on the way they handled the rating task. Data thus obtained not only facilitated the selection of participants for the study but also provided a check on the information provided by participants in the interviews concerning background influences on their behavior.

Interviews. I designed, piloted, and administered interviews following my initial analysis of the verbal protocols, independently of the project that had yielded the data from the protocols and the questionnaires. My aim, besides providing a check on my analysis of raters' behaviors, was to explore raters' backgrounds, and to relate them to the participants' rating behaviors. 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. To counter the risk that the interviewees or I could bias the interview toward socially desirable interpretations of behavior, I analyzed the internal consistency of responses, and compared them with responses in the questionnaires to pinpoint answers that may have had such ulterior motives.

Part 1 of each interview was open-ended. Participants were asked to read excerpts of their concurrent verbal protocols involving all their comments on 12 of the 60 compositions they had rated and commented on in the Cumming et al. (2001) 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 to 5,400 words, a task that took each person, on average, 35 minutes. The sample was made up of participants' comments on two compositions at each of six levels of proficiency in English as assigned by the participants themselves. Whenever possible, the first and last essays rated at each level of proficiency were chosen, to reduce subjective criteria for selection that could influence raters' own recall of the rating sessions. This principle of selection was modified for only the following reasons: Protocols concerning compositions #14, #62, and #128, the only ones that 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; 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 protocols were handed to the participants in ascending order, based on the scores they had assigned to the corresponding compositions: Protocols involving compositions with a score of 1 were given first; protocols of compositions scored 6 were given last, thus reminding participants of the textual features they attended to at each successive level of proficiency. 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 I deemed necessary in view of the amount of time (6 to 7 months) that elapsed between the concurrent think-aloud protocols and the interviews. In addition, by asking participants to comment on the protocols specifically with respect to their backgrounds, I intended to raise their awareness of the range of background factors potentially pertinent to their behavior, in order to prepare for the specific questions subsequently posed in Part 2 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 [their] 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) noted in their related study of raters' backgrounds, "data volunteered by informants may be more valid than data generated by a specific, and perhaps closed-ended, prompt." (p. 241) Including the time spent by participants reading their protocols, I had budgeted 90 minutes for conducting Part 1; the time actually spent on this part eventually ranged between 80 and 110 minutes.

Part 2 of each interview followed immediately after Part 1, and consisted of a question-and-answer session based on my analysis of participants' behavior. Whenever that analysis (described in more detail below) suggested a dimension of variability, I asked each participant to comment on his behavior with reference to his personal background and professional experience. Designed in such a way, the interviews were consistent in exploring an identical range of issues across the four participants (for example, concerning their attitudes to the potential constraints exerted by time pressure and inappropriate topics), but flexible in allowing me to use specific observations concerning participants' behavior in phrasing each question. Participants' responses could also be compared with statements concerning their establishment or revision of scoring criteria in their protocols, as well as with statements they made in the questionnaire concerning background and attitudes.

The purpose of Part 2 was twofold: to get raters to provide comparable data by relating key dimensions of their behavior to their backgrounds and to provide an initial "member check" on my analysis of their behavior. Reflecting my research questions, the interview explored how participants (1) established the meaning of a performance embodied in a composition written under specified constraints; (2) viewed the writing prompts candidates in the sample of compositions were responding to; (3) interpreted the purpose of TOEFL essay; (4) envisaged the candidates writing the test; and (5) constructed scoring criteria. In conclusion, I asked participants whether they considered themselves to be severe graders, with a view to eliciting a summary self-assessment that did not seem redundant in view of the lengthy discussion that had preceded it. I expected Part 2 of the interview to take about 90 minutes, and this estimate proved to be accurate. At the conclusion of each interview, both Part 1 and Part 2 were transcribed in full, though with all references to participants' ethnicity, as well as to educational institutions or professional organizations they had been, or were, associated with, omitted. The transcriptions followed the conventions already established by Cumming et al. (2001) for concurrent think-aloud protocols.

Finally, at the conclusion of my analyses, participants were presented with their own case studies for comments that were designed to act as member checks on my interpretation of the relationship between their backgrounds and their behaviors. This follow-up was entirely open-ended: I simply asked participants whether they felt that my interpretation of their behaviors was accurate, and I invited them to raise specific concerns they had about omissions and inaccuracies.

Data Analysis

Qualitative analysis. My analysis of the concurrent verbal protocols and my construction of the interviews followed in general the precepts of "grounded theory methodology" (Glaser & Strauss, 1967) in that differences between raters emerged from the data rather than being hypothesized prior to analysis. The framework for analyzing the verbal protocols was provided by the research questions that highlighted four broad areas of comparison between raters. Within each of these areas, I searched for contrasts in the behaviors of the four participants, from which dimensions of variability could be extracted. As an example of this procedure, I studied participants' responses to compositions in the corpus that contained a few well-formed sentences and then broke off inexplicably. Chris (as well as Alex) treated such compositions at face value, refusing to compensate for situational factors or to credit someone who so manifestly failed to fulfill the demands of the task for showing command of the language. Sam (and, occasionally, Jo), by contrast, could look at a similarly unfinished essay, speculate extensively on why a candidate broke off after a promising beginning, and give partial credit for showing control of language in the absence of even minimal task fulfillment. Together, these statements suggested a dimension of variability in that participants in the study reacted in different ways to compositions whose authors had clearly failed to fulfill their potential, an issue that relates to the research question concerning raters' views on the significance of an isolated performance embodied in a single composition written under specific conditions.

Whenever I identified such a dimension of variability, I subsequently quizzed participants in the interviews on how they thought their backgrounds influenced their attitudes. My analysis of their responses involved three cycles of interpretive comparisons. Having identified key contrasts in the protocol data, I first looked for confirmation from the participants that such contrasts were indeed present. Then, I sought to identify aspects of their backgrounds that could explain the contrasts by triangulating data from the interviews, the verbal protocols, and the questionnaires. Finally, by making comparisons across the participants, I was able to highlight not only aspects of raters' backgrounds that 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 judgments and behavior.

Quantitative analysis. In addition to qualitative methods in analyzing the protocol data, I incorporated a quantitative component aimed at highlighting the textual features raters attended to during various stages of their rating sessions. This relied on frequency counts of behaviors coded according to a scheme developed in earlier analyses of the same think-aloud data (Cumming et al., 2001). The coding scheme (reproduced in Table 1) was validated by its "precision and relevance" in describing raters' behaviors in assessing a wide range of essays responding to six TOEFL prototype tasks in the Cumming et al. study. It is based on two key distinctions. The first distinction isolates interpretation strategies from judgment strategies and the second distinction, cross-cutting the first, establishes three foci for both types of strategies: self-monitoning, 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., 2001). In adapting it to the present study I added two new codes: "IS0" ("read essay prompt") and "JR0" ("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 behavior labeled as "judge level of difficulty of (or level of challenge posed by) essay prompt."

During phase II of the Cumming et al. (2001) study, another participant and I had used the scheme to independently code, with an average intercoder agreement of 84%, a sample of transcribed protocols that included participants' assessments of 20 of the 144 compositions at various levels of proficiency. For the present study, I recoded the full transcripts of the rating

Table 1

Coding Scheme of Decisions Made While Rating ESL Compositions (based on Cumming et al., 2001)

Self-monitoring focus	*Rhetorical-ideational focus*	*Language focus*
Interpretation strategies		
IS0 read essay prompt		
IS1 judge level of difficulty of (or level of challenge posed by) essay prompt	**IR1** interpret ambiguous or unclear phrases	**IL1** observe layout
IS2 read/reread composition	**IR2** discern rhetorical structure	**IL2** classify errors into types
IS3 envision personal situation of writer	**IR3** summarize ideas or propositions	**IL3** “edit” phrases for interpretation
IS4 scan composition		
Judgment strategies		
	JR0 consider writer’s use and/ or understanding of prompt	
JS1 decide on macro-strategy for reading and rating (either for a paper or for a corpus)	**JR1** assess reasoning, logic, or topic development	**JL1** assess quantity of total written production
JS2 consider own personal response or biases	**JR2** assess task completion	**JL2** assess comprehensibility
JS3 define and/or revise own criteria	**JR3** assess relevance	**JL3** consider gravity of errors
JS4 compare with other compositions or “anchors”	**JR4** assess coherence	**JL4** consider error frequency
JS5 summarize, distinguish, and tally judgments collectively	**JR5** assess interest, originality, creativity, sophistication	**JL5** assess fluency
JS6 articulate general impression	**JR6** identify redundancies	**JL6** consider lexis
JS7 articulate or revise scoring decision	**JR7** assess text organization	**JL7** consider syntax and morphology
	JR8 assess style, register, or genre	**JL8** consider spelling and punctuation
	JR9 rate ideas and rhetoric	**JL9** rate language overall

sessions of the four raters whom I had requested to participate. Then, having broken down each participant's rating session into distinct stages, I tabulated the total number of coded comments with self-monitoring, rhetorical-ideational, and linguistic foci for interpretation strategies as well as for judgment 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 judgment strategies for the different phases of participants' rating sessions.

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. The data revealed, for example, that "comparing compositions to other compositions or anchors" represented 7.7% of all coded behaviors in Jo's protocols and 3.8% of all coded behaviors in Alex's protocols but occurred not once in either Sam's or Chris's protocols. Combined with the fact that (as discussed earlier) Alex and Jo each read every composition at least twice, and sometimes up to four times, whereas Sam and Chris read them only once, significant differences emerged in the reading and rating strategies of the four participants (the subject of the third research question), and these could once again be explored in the interviews. Interestingly, the responses also revealed that while Jo and Alex followed the procedures they had established in rating large corpora of compositions in standardized tests, and Sam followed the procedures he had developed in rating student essays in classroom assessment, Chris departed from his normal practice in response to what he perceived to be the purpose of his assessment, showing the important (and unpredictable!) role played by specific instructions given to raters at the outset of a rating session:

> I think [my behavior] was related to the project itself, that the instruction was to rate [the essays] 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. I knew that because of the nature of the project what was, what would be looked at was not, in fact, the number at the end but rather the process and the comments that I had made during my think alouds, so I didn't really see much of a value. And, at a very practical level, I don't know if I had time ... 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, I just didn't bother.

Coding of the protocols also supported the identification, initially based on my impressionistic readings, of distinct stages in each participant's reading of individual compositions. This was important in specifying participants' scoring criteria, the subject of the fourth research question, because their comments at different stages in their rating sessions had different implications. At the same time, largely because of the small number of participants (17) for whom frequency counts were compiled, many of the differences in frequency counts of coded behaviors lack statistical significance, meaning that the analysis was much more qualitative than quantitative in its general thrust.

Results

Construction of the Case Studies

Following the initial phases of analysis, I constructed four case studies to highlight key contrasts between the way the participants in the study performed the rating task. The case studies are comparable in that each addresses the same sources of variability, in the same order, using identical sources of data and identical methods of analysis. Table 2 lists the key sources of variability pertinent to each of my research questions as they emerged from the analysis, and the relevant data, providing not only a summary of the research methods followed in this study, but also a road map for reading the study's results, presented below, as well as a starting point for future studies.

Research Question A: What reading strategies did the raters of ESL compositions participating in the study establish to deal with both individual compositions and a corpus of compositions, and how did background factors influence these?

Because scoring procedures provided the context for statements concerning compositions in the concurrent verbal protocols, the strategies participants followed in reading both individual compositions as well as a corpus needed to be laid bare prior to examining the construction of scoring criteria. Briefly, two distinct strategies could be observed: Sam and Chris followed a read-through approach (cf. Milanovic et al., 1995), reading a composition only once and giving a score immediately afterward, whereas Alex and Jo first sorted compositions into preliminary

Table 2

The Structure of the Case Studies

Sources of variability	Sources of data
Research Question A	
– strategies for reading individual compositions	– impressionistic comparisons of sequences of coded behaviors within participants' protocols dealing with individual compositions – participants' comments in the interviews concerning their reading of individual compositions
– strategies for reading a corpus of compositions	– impressionistic comparisons of sequences of coded behaviors within participants' rating sessions as a whole – participants' comments in the protocols and the interviews concerning their strategies for reading a corpus of compositions
Research Question B	
– relating performance to proficiency	– general statements relating performance to proficiency – statements relating the mastery of specific competencies to proficiency
– attitudes to situational factors affecting writers' performance	– statements concerning the nature of language proficiency and language learning – scores assigned to unfinished essays and participants' comments about such scores – statements inferring writers' potential from their performance
Research Question C	
– views on biases in writing prompts	– frequency of references in the protocols to test takers' ethno-linguistic or sociocultural backgrounds – explicit statements concerning biases in the writing prompts – participants' assumptions concerning test takers – participants' assumptions concerning test use
– views on the role of writing prompts	– participants' views of the cognitive demands set by the prompts – frequency of references in the protocols to task fulfillment or relevance – participants' assumptions concerning test takers – participants' assumptions concerning test use

(table continues)

Table 2 (continued)

Research Question D	
– views concerning a developmental trajectory for language learning	– comparisons of participants' comments linking proficiency levels to the mastery of specific competencies (in the protocols or the interviews) – comparisons of proficiency levels established by participants for the lowest and highest points in their rating scale (in the protocols or the interviews)
– specific scoring and criteria	– proportional frequencies of comments with a linguistic focus with a rhetorical-ideational focus in participants' protocols – comparisons of participants' statements concerning relative importance of linguistic and rhetorical-ideational features (in the interviews) – comparisons of proportional frequencies of comments concerning specific textual qualities in participants' protocols – comparisons of participants' statements establishing criteria for specific scores (in the protocols)
– external criteria	– comparisons of pass/fail thresholds established by participants (in the protocols or the interviews) – comparisons of specific performance expectations (as discussed under Research Question B) – comparisons of external norms (and reference populations) referred to by participants (in the protocols or the interviews)

piles, and then read the compositions in each pile up to four times before deciding on final scores. In spite of these differences, however, and in spite of further contrasts between Alex, who adopted the "principled two-scan read" approach (ibid.), and Jo, whose procedure was more elaborate, three basic stages were evident in participants' protocols concerning individual compositions: First they (usually, though not invariably) scanned a composition for length and appearance; next, they read the compositions and provided a running commentary, during which they assessed certain textual features; then, having finished their reading, they distilled their impressions in a final summing up, and assigned a score. As an example of this procedure, consider Sam's reading of essay # 132:

> **(stage 1)** Next one, 132, a long one, typed, over half a page. **(stage 2)** 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. Uh, 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 the author meant. Oh, bad, bad sentence structures. Problems with the plural of the nouns. Problem with the verb tenses, uh, subject-verb agreement. Uh, really bad problems with verbs, person and tense. And singular and plural. **(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.

Because differences could be observed in the textual qualities participants commented on during the different stages of their rating, which frequency counts of coded behaviors confirmed, it was important to identify these qualities within participants' reading of individual compositions and within their rating sessions as a whole. To give just one striking example, during the running commentary stage of Chris's individual protocols, the proportions of interpretive and judgmental strategies with a rhetorical-ideational focus were, respectively, 6% and 67%, and the proportions of interpretive and judgmental strategies with a language focus were, respectively, 23% and 30%. During the summing up stage, the proportions of interpretive and judgmental strategies with a rhetorical-ideational focus rose to 37% and 52%, respectively, and the proportions of interpretive and judgmental strategies with a language focus declined to 2% and 19%, respectively. The pattern emerging from these data, that Chris frequently commented on language control but seldom took it into consideration in assigning a score, was fully confirmed in the interview:

> 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 light of such information, there was an obvious need to analyze participants' strategies for dealing with the compositions. However, although the analysis yielded much interesting information, its only relevance to the present study was in specifying the contexts in which participants' statements had to be placed before being interpreted. The only other pattern worth noting was that assessment experience played the key role in shaping the strategies of all four participants. Sam used the procedures he had developed in classroom assessment in his native country, where he would read each composition only once before giving a score; he justified this procedure in the present context by referring to time constraints and by expressing a high level of confidence in his judgments. Alex and Jo followed the elaborate procedures they had found useful in dealing with large corpora of compositions in their experience with standardized assessment. Only Chris departed from the procedures he normally followed in dealing with placement tests (the closest that his experience came to the anonymous setting of large-scale, standardized assessment) by not sorting and rereading compositions; he attributed this to the lack of practical consequences that his grades carried in what was an experimental, rather than an assessment, context. Such behavior exhibited by raters in the present study suggests that rating strategies, particularly macrostrategies employed in rating a corpus, may represent an aspect of raters' behavior that can be fine-tuned through focused training and practical assessment experience, or, in Pula and Huot's (1993) terms, through participation in an "immediate discourse community." (p. 256)

Research Question B: In what principled way did the raters 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?

Participants' attempts to infer a writer's proficiency from her performance were evident from the verbal protocols, particularly when the scores assigned to compositions were linked to the mastery of specific competencies:

> [ESSAY #70] Short one, five lines typed. Well, I guess, 1 or 2 will be the grade. Let me read it first. [...] OK, I'm reading it ... some statements there, uh, language, bad grammar. There is a whole sentence which is OK: [QUOTE] Another correct sentence: [QUOTE] And because there is full sentence structure I would give it a 2, not a 1. So 70 gets 2, it's bad but it's not the worst. /Sam/

> Now, script 44. [...] Actually very few errors. It's short. So, it's not rich in content, and the organization is not good, because it's just one chunk of words. However, it does, it does satisfy the minimum requirement for the task [...]. There is no leading sentence or leading paragraphs, nor is there a conclusion, but it has the –, it it should be a 3 at least and it's very likely that this writer has good proficiency. So, in view of that, I may just put a 4 down because I understand that I'm based on this to infer as to the writer's proficiency. I think he's a good writer, uh, judging from the proficiency, from the language. /Alex/

Subsequently, when asked in the interviews to comment on the criteria used in assigning scores, participants were even more explicit in associating the mastery (or lack of mastery) of specific competencies with a writer's level of development. This suggested that they linked performance to proficiency through internalizing a developmental trajectory for ESL learners, so that the scores they assigned represented, in effect, their opinions of where writers stood on that trajectory. For example, Chris, who in his concurrent verbal protocols appeared to be the least conscious of following such a procedure, considered a writer who repeated an essay prompt in the opening sentence of her composition to have only limited proficiency:

> Again, from my teaching experience, I find that is a strategy often used by, when 1 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.

Conversely, he considered creativity (manifested, for example, in the mastery of a range of sentence structures) to be a sure sign of advanced proficiency:

> If it's something interesting, somewhat creative, [...] it shows me that the writer is able to manipulate the language in a way, to get beyond just very simple basic sentence structures and stating things, and it to me that shows a higher level of writing ability. So if I see that, if I see evidence of that, I think, oh this is really, this writer is really developing, is coming along quite well, if he or she is able to do that.

Key textual characteristics that the other participants associated with proficiency levels in the interviews included, for Alex, degree of flexibility in paragraph structure and cohesive devices; for Jo, length of response and degree of detachment; and, for Sam, handwriting, mastery of adverbs, rigid paragraph structure, and length.

Such associations, which recall the view that readers' judgments of graduate students' compositions are at least partly based on inferences about writers' efforts, thought, and knowledge (Prior 1995, p. 53), suggested that the four participants in the study shared the goal of inferring proficiency from performance. However, the interviews also revealed that participants differed in their definitions of language proficiency and language learning. Such differences, in turn, carried over to participants' attitudes toward the impact of situational factors, such as topic effects or time pressure, on performance, as discussed below, and contributed significantly to differences in participants' scoring criteria and, consequently, to variability in scores participants assigned to compositions, as discussed under Research Question D.

Alone among the participants, Sam took the view that language proficiency (and, more specifically, writing proficiency) reflected a person's intellectual development in general:

> I don't think 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.

Looking at the influence of background factors, his attitudes were determined principally by his learning and teaching experiences prior to his coming to North America. Because he felt that it was the curriculum followed in his native country that had enabled him, his peers, and his students to successfully acquire English, he was confident in relying on the educational concepts that had produced that curriculum. These concepts centered on the notion that language proficiency was a measure of intellectual ability in general, something that Sam also deduced from his personal experiences learning English and writing language tests, and that led him at times to label writers as "intelligent" (#26), "confident" (#40), "unintelligent," (#10, #62), "immature" (#113), or "not serious" (#14) in the concurrent verbal protocols.

Beyond leading Sam to pay close attention to grammatical competence, these concepts generated few specific scoring criteria. However, relying on such a far-reaching definition of proficiency, Sam could not only identify a writer's position on a developmental trajectory, but also compensate for situational factors that may have negatively affected performance. This is exemplified by his reaction to essay #126, which broke off after a few well-formed sentences:

> That's the strangest thing I read until now. Uh, it has an introduction, and uh, it has good language, it has good structure in these ... three sentences. This is definitely a beginning of a good essay, but it's just a beginning, I have no idea why this person didn't finish. What's written here is very good, it's a 5. Even if there is some poor word choices and some grammar problems. Uh, I'd say this person, this person just either became sick or I don't know whether it is possible but can you be cheating on such a test, and can you sneak behind someone and over the shoulder write what another person has written. I don't know, but this is very puzzling. Anyway I have to give it a grade, uh. I'd give it a 2 because, uh, there is no essay whatsoever, but, uh, it's evident that the person has good command – I'm sorry, I didn't realize the tape – I guess what I was trying to say was that,

> uh, it shows some command of the language, it shows fluency, so, I'd say I'd give it a 2 because – not because it is an essay, but, uh, because of the good language. /Sam/

Indeed, Sam at times went a step further by not only estimating a writer's level of proficiency, but, at times (as in the case of the writer of essay #133), even inferring her potential. In his interviews, he attributed his ability to thus compensate for underachievement partly to teaching experience, and partly to his strong focus on grammatical competence (itself a product of the educational system that had nurtured him as both learner and teacher), which to him was the clearest index of proficiency:

> 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.

Unlike Sam, Chris initially arrived at the idea of placing writers on a trajectory through his experience with placement testing, where he had to fit learners into a curriculum for writing instruction based on a single performance:

> 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 for each level.

In subsequently developing a sense of what competencies could be mastered by writers at successive levels of proficiency, Chris relied on generalizations of how writers whose overall proficiency he could ascertain through continuous classroom observation performed on tests, as well as on his understanding, and acceptance, of the concepts of communicative competence and communicative language teaching. And, because the communicative approach does not associate language ability with intelligence, Chris's views of language proficiency and language learning were at once more restricted and more holistic than Sam's: On the one hand, he held language ability to be independent of mental abilities in other spheres; on the other hand, he did not see grammatical competence playing an overwhelming role in instruction as it was bound to emerge through writers' continuing efforts to communicate their ideas:

> I think attention to grammar is important, and, and I think that it is something that can be taught and learned, ... [BUT] uh, 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. [...] 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 addition, Chris did not take situational factors into account in the process of rating compositions, regarding them only as complicating factors in an already complex process. His ideas in this area emerged most clearly from his reactions to an unfinished composition:

> 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 TWE. So, anyway, a score of 1; there is nothing here.

He likewise refused to assess a writer's future potential, restricting his assessment to inferring a writer's current level of development from her performance. His 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.

As for Alex, he alone identified "understanding of the relationship between proficiency and performance" in the questionnaire as a key factor influencing his assessment of compositions. Although he did not elaborate, he gave glimpses of his attitude in the concurrent verbal protocols and 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 [THOUGH] 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. I have, actually I taught graduation level for many years, and I rated the same public exam for many years. So, that kind of an outside-testing-context knowledge could help me associate a certain performance with a certain level of proficiency.

In particular, his experience of teaching ESL at both secondary and tertiary levels in his native country led Alex to equate language ability with communicative competence. At the same time, perhaps because he had been teaching English in an EFL, rather than ESL, context, he held that grammatical competence could be improved through focused instruction and not merely through attempts to communicate. He also viewed grammatical competence as a prerequisite to effective communication, though not as its final goal, which, along with his views of the purpose and limitations of TOEFL essays, may explain why he could give credit for language proficiency on the one hand, but do so mostly at the low and middling levels of proficiency, on the other:

> ... my question was "Whether or 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?" ... for this project, because I know it's 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 the 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. [Alex's assessment of essay #44, cited earlier, providing a case in point]

Apart from teaching experience, Alex's 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, and to focus on grammatical competence. However, he was aware of the criticism directed at this hypothesis and did not derive any specific scoring criteria from it, beyond stating that the order in which syntactic and morphological structures were acquired could constitute a developmental trajectory:

> 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 and actually has formed 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 overstated. I mean I think that the use of various information in rating may have sometimes come from the theory in the literature. And, although this kind of research has been challenged, I think there is some kind of gradation there in the acquisition of certain grammatical structures.

As for situational factors, particularly time pressure, Alex could appreciate their impact on a writer's performance, but made no allowances for them. This 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, 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.” This, in light of his assertion that in performance-based assessment one tries to estimate proficiency from an isolated performance, must mean that by “rating a performance at face value” Alex was trying to infer proficiency without speculating what a writer might have been able to do under ideal conditions.

The fourth participant, Jo, associated performance with proficiency as a result of experiences with standardized assessment as well as test development, including the piloting of writing prompts and the establishment of scoring rubrics based on writers’ responses. These experiences had given him an idea of what kind of performance writers at various levels of development were capable of, and led him to view scoring rubrics as reflective of a writer’s developmental trajectory (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.”

Note, however, that having derived broad scoring 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 Cumming et al. (2001). In addition, although he did not relate language proficiency to intellectual development, Jo's experience with test development and standardized assessment also 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 affect performance, especially when the stakes are high:

> [ESSAY #95] Looks like [THE AUTHOR] 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.

In sum, because inferring proficiency from an isolated performance lies at the heart of performance-based writing assessment, understanding the procedures raters adopt to make such inferences is crucial to validating tests where candidates are rated on compositions they have written under carefully specified constraints. In other words (Bachman 1990), "the distinction between language ability and the performance of that ability [is] at the same time a central axiom and a dilemma for language testing." (p. 308) What united the four participants in the present study was that they all attempted to bridge the gap between performance and proficiency (or, in Bachman's terms, the performance of language ability and language ability itself) by internalizing a developmental trajectory for ESL learners, by equating proficiency with a position on that internalized trajectory, by looking for features in each composition that could identify that position, and then by expressing that position as a score. Conversely, what divided the participants was that they differed in their definitions of language proficiency, internalized different trajectories of language learning, and differed in the range of information they collected. The implications of these findings, beyond differences, already noted, in participants' attitudes to the impact of situational factors on performance, will be discussed in answering Research Question D, dealing with the establishment of specific scoring criteria.

Research Question C: How did the raters participating in the study interpret the role of writing prompts in performance-based assessment, and how did background factors influence this process?

Two related sources of variability emerged from my analysis of the data concerning attitudes to writing prompts: the degree to which participants considered essay topics to be discriminatory and the importance participants attached to relevance and task fulfillment in writers' answers. Regarding the first issue, Kroll (1998, p. 223, citing Hamp-Lyons & Kroll, 1997, p. 21) had 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," and my data show that participants clearly differed in their assessments of the cultural and socioeconomic gulf separating writers and essay prompts.

Sam, in particular, was not averse to criticizing prompts that placed unrealistic expectations on disadvantaged candidates, as shown in his comments in the protocol concerning essay #24, responding to a prompt asking candidates to identify the most important room in a house: He obviously felt that it was unfair to ask candidates about rooms in a house, because many of them may not have ever lived in a house. He also showed his sensitivity to writers' personal backgrounds by making inferences concerning their gender (in essays #20, #21) and socioeconomic backgrounds (in essay #24), and by noting references made by writers themselves to their backgrounds (in essays #18, #40, #128).

Alex, likewise, was aware of the potential shortcomings of the essay topics used in the present study. However, although at times he tried to infer a writer's mother tongue (#22, #48), academic orientation (#74, #91), or age (#5), such comments came in the form of asides during his initial scanning of the compositions and at one point, reading essay #21, he even cautioned himself against speculating about writers' backgrounds. He also acknowledged that in the interests of comparability the topics on a standardized, international test such as TOEFL had to be fairly universal even at the cost of handicapping writers from certain cultural backgrounds:

> This sort of, uh, prompt [ASKING CANDIDATES IF THEIR HOMETOWNS WERE SUITABLE LOCATIONS FOR A NEW UNIVERSITY], I mean they are not ideal, [...] they are there, because, it's almost like inevitable. You want a question or a task that can be applied to different cultures, to students of the world, you know, who want to come to North America to study. But that student may be from Iran, Morocco, Japan, etc., so you cannot make it a very contextualized prompt. It would have to be something very hypothetical or universal. So, sometimes, uh, you try to contextualize it but at the same time you know that it's not really, it's a universal thing.

By contrast, Chris and Jo found the essay topics to be unproblematic. Chris did read each of the four topics at the outset of his rating session, "to get an idea of what the tasks were," and he confirmed during the interview that he did so to detect biases; however, he simply did not find any bias worth noting. In addition, although he commented on geographical names in compositions when they referred to places familiar to him (#21, #37), he not only never speculated about writers' ethnicity, language, or culture, but consistently overlooked explicit references to such background factors by the writers themselves (#40, #128). As for Jo, he explained in the course of the interview that he had assumed that the writing prompts for such a widely used test as TOEFL would have already been vetted by test developers. Consequently, although at the beginning of the rating session he analyzed the essay prompts from the point of view of accessibility to candidates, his principal concern was with the level of challenge, as well as level of difficulty, posed by the writing prompts. In addition, like Chris, Jo resolutely avoided any speculation concerning writers' ethnolinguistic or socioeconomic backgrounds, even when such information was offered by the writers themselves. However, Jo did compensate for topic effects to the extent that he derived his rating scale during his rating session at least in part from the range of responses found in the corpus of compositions he was actually rating, instead of importing all his scoring criteria from another context, such as his own work with test and rating-scale development.

Although it is impossible to generalize from a small sample, the fact that the two NSs found the essays to be neutral while the two NNSs found them to be biased is noteworthy; apparently, raters' attitudes to biases in essay prompts are among the few areas of variability that may, indeed, be explained with reference to a simple contrast between NSs and NNSs of English. What is more important, however, is that none of the four participants compensated for biases in essay prompts, not even the participants who clearly identified the presence of such biases.

Sam showed in his assessment of the writing prompts at the outset of his rating session that he viewed topics principally as vehicles that permitted writers to express complex ideas: He found Topic A (asking candidates to choose the most important room in a house) and, particularly, Topic D, which forced writers to choose between two simple alternatives (one long versus several short school vacations), "very limiting." By contrast, he approved of Topic B (asking candidates to decide if their community would be a good place for locating a university) and Topic C (asking candidates about the relative importance of certain academic subjects), because they allowed for good argumentation. Further, Sam explained in the interview that his speculations in the protocols concerning writers' cultural or linguistic backgrounds arose out of a personal curiosity, rather than out of a desire to compensate for discrimination. Finally, referring to his own experience as a test taker, he expected good writers to cope even with socioeconomically biased essay topics, by using their critical faculties or their linguistic competence:

> 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 [...] 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 higher grades.

Conversely, because in Sam's view writers commanded numerous options to demonstrate their intellectual development (which is what writing tests really measured for him), he felt that raters of ESL compositions had a wide range of criteria to aid their assessment, rendering overt compensation for discriminatory topics unnecessary.

Alex likewise used his experience with assessment, although from the vantage point of the assessor not the candidate, to argue that good writers could cope with poor topics. Although be did tailor his scoring criteria to reward such coping strategies as the adoption of a persona, he, like Sam, found overt compensation for biased essay topics unnecessary:

> So, [THE WRITER] 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.

From a similar vantage point, Chris felt that good writers could cope with poor topics by modifying them to suit their personal situations:

> I didn't feel having to write on [WHAT IS THE MOST IMPORTANT ROOM IN A HOUSE] was culturally biased, or insensitive. I think students read into a topic and they apply it to their situation automatically anyway.

Finally, Jo did not even rely on generalized observations of how good writers behaved, but simply concluded early on in his rating session for Cumming et al. (2001) that because certain candidates (specifically, 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.

In explaining their reluctance to compensate for topic effects, participants relied on two specific arguments. The first was their desire to avoid “second-guessing” or becoming attached to certain writers:

> 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. /Jo/

> 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. /Chris/

Evident in most of these comments is a distinction between teaching, where raters should keep writers' backgrounds, personalities, and abilities in mind to enable personalized instruction, and assessment, where such factors must be ignored. Less evident from the comments, but equally sharp, are distinctions between classroom tests and standardized assessment: In any single classroom test a rater can avoid inferring a writer's level of proficiency, which he may more reliably estimate through continuous observation, and focus strictly on her performance; in standardized assessment, however, proficiency is precisely what a rater must estimate (usually by matching the observed performance with rating scale descriptors, which may themselves reflect assumptions about learning a (second) language (cf. Bachman & Cohen, 1998). It is significant that whereas all four participants in the study adhered to these contrasting approaches, only the two with extensive experience with standardized assessment (Jo and Alex) appeared to be fully aware of doing so. Sam, in particular, was torn between theory and practice: He felt that inferring a writer's level of proficiency from performance was a "subjective" procedure that should be avoided, but he repeatedly performed the procedure nevertheless.

More importantly, even if not all participants fully appreciated the meaning of "objectivity" in performance-based assessment, they were united in their view that the purpose of TOEFL was to screen potential university applicants for their ability to cope with the demands of studying in a foreign language, and formed clear expectations from the writers of the compositions they were rating. Although these expectations varied, they outweighed participants' interest in writers' personal backgrounds and led participants to view topics as challenges that writers aiming to be university students had to be able to overcome. This 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. Whatever the justification, they all had the effect of rendering overt compensation for topic effects unnecessary. What the results of the study seem to indicate here, then, is that differences in raters' backgrounds have little impact on how raters view writers and writing prompts, if they agree broadly on the purpose of a test and, consequently, form similar expectations of the candidates writing it.

A related issue that emerged subsequent to analyzing the verbal protocols was that two of the raters regarded the essay topics as tasks that had to be fulfilled or, even, as questions that demanded answers. Alex, in particular, consistently assessed task fulfillment, and frequently equated it with "answering the question." This corroborates the evidence of the questionnaire, where he stated that he regarded task completion as the key to success in composition examinations (although he was willing to overlook a degree of irrelevance if the writer's language was otherwise proficient, as in the case of essay #4). Chris, likewise, referred to essay prompts as "questions" that demanded answers, and "answering the question," that is, addressing the topic and taking a clear stance were important criteria in assessing the compositions in his corpus (with essays #140 and #3 providing examples of, respectively, effective and ineffective answers).

Sam and Jo, on the other hand, provided no evidence either in the protocols or in the questionnaire that they were looking for answers to questions. A plausible conclusion is that they viewed the writing prompts merely as vehicles for eliciting a performance. This would explain not only their relative lack of interest in task fulfillment, but also their attitude, discussed above, that the more challenging a prompt and the more complex the response it permitted, the better. This attitude allowed them not only to overlook occasional misinterpretations of the prompt on the part of writers, but also to refuse to compensate writers they themselves may have identified as disadvantaged: Because prompts were merely vehicles for eliciting language performance, any disadvantages a writer may have suffered from should have been surmounted through appropriate strategies and, above all, linguistic competence.

As with their views concerning biases, participants referred almost exclusively to their encounters with assessment in explaining their attitudes to the role of writing tasks. 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) about how to precisely interpret a topic (See Connor & Carrell, 1993, for similar findings). Like Sam, he thus treated writing tasks as prompts as long as writers were not merely avoiding a topic in order to recite memorized bits of language. In contrast, Alex and Chris consistently expected writers to fulfill the requirements set by a task.

Alex did so because as a rater he had found the ability to fulfill the specific demands of a task to be a good discriminator among relatively proficient students (and because he had been accustomed to task-based teaching at higher institutions of learning in his native country). Chris, on the other hand, emphasized that in his experience with both classroom assessment and placement tests he had seen too many perfect off-topic essays that were clearly memorized in the hope of getting a high grade.

Taken together with participants' attitudes to biases in writing tasks, it appears that Sam and Jo had developed a greater sense of empathy to writers in general. This suggests that it is not the level of assessment experience that is the determining factor here, but the nature of raters' 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 so much 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.

Research Question D: What specific scoring criteria did the raters participating in the study generate in the absence of a scoring rubric, what information did they heed during this process, and how did background factors influence their choices?

As discussed under Research Question B, in the absence of a scoring rubric the initial step for all four participants in the establishment of scoring criteria involved making the assumption that they could construct a developmental trajectory for ESL writers and identify specific criteria to indicate a writer's position on that trajectory. After taking that initial step, participants made assumptions about test takers and test use that, as discussed under Research Question C, determined the way they viewed the role of writing tasks in performance-based assessment. What the fourth research question explores is the way participants generated specific scoring criteria from their general assumptions concerning language proficiency and language learning, including what additional information they considered, what specific criteria they established, and how their backgrounds influenced this process.

Sam began with the assumption — arising out of mutually reinforcing experiences as learner and teacher in the same educational system, and out of social attitudes to language in his native country — that linguistic development was a sign of intellectual development in general. This led him, above all, to place a premium on correctness, which was not only a prerequisite for the acquisition of other competencies, such as fluency, and the mastery of registers and idiomatic expressions, but also a sure sign of intelligence:

> 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 leaning 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.

The second important sign of intellectual development for Sam was the ability to generate ideas and organize them in a logical manner. It is at this point that his assumptions about the purpose of TOEFL (as a test meant to screen potential university students for their English proficiency) came into play, because, in his opinion, students lacking the ability to gather and organize ideas would have a hard time succeeding at a 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. Hmm, 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 a consequence of his views on language proficiency and the purpose of TOEFL, Sam also held that the principal role of the writing task was to elicit a language performance. These assumptions helped him set additional performance expectations: to wit, that writers should be able to respond to any of the four topics used in the administrations of the test that had yielded the compositions and, at the same time, that strict adherence to the topic in a response was secondary to the demonstration of language control.

As a result of these initial assumptions, it is not surprising that comments accompanying the scores Sam assigned to compositions focused heavily on “syntax and morphology,” “overall language ability,” “reasoning, logic, or topic development,” “length,” and “text organization.” Proportional frequencies of coded behaviors in his protocol also revealed that he commented on language more frequently than on rhetoric and ideas, particularly in phase III of his protocols (when he assigned scores): In that phase 6% of his interpretive and 37% of his judgmental strategies had a rhetorical-ideational focus, and 32% of his interpretive and 54% of his judgmental strategies had a rhetorical-ideational focus. At the same time, he shifted focus slightly from language to ideas and rhetoric as he assigned successively higher scores to compositions (cf. Pollitt & Murray, 1995) — with the exception of compositions he rated at the top level, whose language appears to have so impressed him that he set aside his interest in ideas and rhetoric.

In generating the specific criteria that allowed him to assign scores, Sam also took into account the instruction that he was to use a 6-point scale, and began by defining the endpoints. He assigned a score of 1 to compositions that demonstrated no ability to string either ideas or words together. Conversely, using his own performance on the Test of Written English as a yardstick, he assigned a 6 to compositions that corresponded to his image of a competent NNS writer. Then, as confirmed in his interview, Sam parceled out his scale by initially categorizing essays as “good,” “middling,” or “bad,” and expanded his scale to six points by finding criteria to separate the three broader levels. Finally, he pegged a passing score at 5, based on his information concerning passing scores on TOEFL essays, and to meet this score, writing had to meet his expectations for prospective university students.

As for specific criteria generated through this procedure, an essay that Sam rated at 1 exhibited, in his view, no organization, no development of ideas, and no mastery of sentence structure. Essays he rated at 2 again lacked any development of ideas, but showed evidence of either sentence-level mastery (e.g., #11, #62) or rudimentary organization (e.g., #15, #137). In essays Sam rated at 3 argumentation emerged, most requisite structural elements (such as topic sentences) were present, and paragraphing was mastered; however, language control remained poor (e.g., #117, #37). Essays rated at 4 were similar to essays rated at 3 except that they showed improved control of language and/or evidence of original thought (e.g., #98). Similarly, essays Sam rated at 5 exhibited fully developed arguments but lacked a clear mastery of the language, something that, along with creativity, only essays rated at 6 possessed. As this scheme suggests, Sam's interim scale of three broad levels of proficiency (1-2, 3-4, and 5-6) was determined by assessing both argumentation and language control. However, because Sam basically took the quantity and quality of argumentation to represent a writer's performance, and language control to reflect her proficiency, it is not surprising that in assigning a final grade on the 6-point scale, Sam accorded greater weight to language control.

In explaining the influence of background factors on his scoring criteria, Sam repeatedly stressed his learning experiences in his native country, reinforced by his teaching experiences in the same educational system. That system focused heavily on grammar, partly because it was felt to be teachable (and learnable), partly because, as Sam explained, its mastery, in his experience, enabled students to acquire the other elements of communicative competence, and partly because the method of focusing on grammar worked in practice. By contrast, the alternative method of immersing students in English and letting grammar emerge out of their efforts to communicate was considered neither feasible in a country where English was spoken only by a select few, nor acceptable. Sam himself was most emphatic in his protocols that “you cannot learn by osmosis” and that fluency, unlike grammatical competence, was easy to achieve through exposure, but did not guarantee success at the university level.

Embedded in Sam’s views are cultural attitudes to education, including the assumption that if, for example, grammar is teachable and learnable, then one’s level of grammar is indicative of one’s level of schooling and one’s level of intelligence (which, according to this set of assumptions, amount to almost the same thing). Cultural background also showed up in Sam’s appreciation of creativity, a key requirement for getting a score of 6: Sam felt that prizing creativity was a form of opposition to the enforced conformity he had experienced in the political system of his native country. Finally, cultural attitudes also emerged in Sam’s dislike of colloquial expressions, which recalled for him the politically motivated destruction of academic standards in his native country, and in Sam’s awareness of a contrast between approaches to rhetoric and argumentation in his native country and in North America.

However, cultural assumptions only worked for Sam if his practical teaching or learning experiences confirmed them: As seen above, he advocated the teaching of grammar not only for theoretical reasons but also for the practical one that it actually worked for him, his peers, and his students in his native country. Likewise, he may have originally prized creativity for cultural reasons, but he also noted that his creative approach to compositions repeatedly earned him top grades on standardized assessments in North America even though he followed a different rhetorical style. Finally, as shown by his definition of adequate topic development, Sam clearly placed practical experience above theoretical knowledge, although this may have been done out of expediency rather than conviction:

> 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.

Chris's scoring rubric was rooted partly in the principles of communicative language teaching, which led him to assert, above all, that there was little point in teaching grammatical structures directly, because these “worked themselves out” in the effort to communicate. It is important to stress that he did not thus downplay the importance of language control, but considered it a relevant criterion mostly for writers he considered proficient. Other general expectations, for Chris, arose out of teaching experiences; for example, he advocated formulaic essay organization for lower level writers as a means of adequately structuring their ideas and, conversely, expected creativity and an academic style from writers near the top of his developmental trajectory.

In generating more specific scoring criteria from these general expectations, Chris, like all other participants, made use of the information that he was to create a 6-point scale. To this end, he first read several compositions that contained only a few sentences and lacked evidence of paragraphing, and thus established the bottom end of his scale. Then, he took the construct of an educated NS as the norm for assigning a 6 to compositions, because the rating scales he was familiar with had all done so. However, unlike all other participants, Chris did not set a pass/fail threshold for this rating task: Although he was aware that the purpose of TOEFL was, normally, to screen prospective university applicants based on their English proficiency, he also knew that his ratings in the present context were undertaken for research purposes and carried no practical consequences.

Proportional frequencies of Chris's coded behaviors demonstrate that he focused on rhetorical-ideational qualities to a much greater extent than on language throughout his ratings, particularly when making judgments during the final phase of his reading of individual compositions: In phase III of his protocols, 37% of his interpretive and 52% of his judgmental strategies had a rhetorical-ideational focus, but only 2% of his interpretive and 19% of his judgmental strategies had a language focus. He placed especially strong emphasis on "reasoning, logic, or topic development" and he not only assessed these traits more often than any other, but also placed essays roughly in two groups along this dimension: Compositions he rated at 3 or below received overwhelmingly negative comments concerning "reasoning, logic, and topic development," and those he rated at 4 or 5 received mostly positive comments. Additionally, Chris looked for the key structural elements of an essay: "adequate" topic development (including both a certain quantity of information and a degree of coherence); a clear stance; "academic" tone, defined largely negatively as the absence of such dramatics 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. Overall, essays that were incomprehensible, irrelevant, or simply too short to allow topic development got a 1; essays showing development, but not addressing the topic were awarded a 2, even if they were quite long or fairly accurate in their language; essays showing topic development, but illogical, incoherent, or not fully addressing the topic got a 3, even if they were creative and linguistically adequate; essays that were well organized, on topic, and developed, but boring, error prone, or occasionally unclear got a 4; and essays that were well organized, on topic, and developed but contained minor linguistic and organizational flaws got a 5.

Regarding the origins of Chris's scoring criteria, his comments on his background provide the clearest support in my data for Pula and Huot's (1993) thesis that enculturation in discourse communities is a useful model for explaining the behavior of raters, although, unlike raters in Pula and Huot's study, Chris relied much more on his teaching than on his assessment experience for scoring criteria. This teaching took place in an educational institution that favored communicative competence over a narrow focus on grammar and advocated a process approach to teaching writing, assumptions that were incorporated into the curriculum and also determined the informal, in-house rating scale that Chris used in placement testing. However, just as Sam's theoretical views on language proficiency and language learning were reinforced by practical experience, so Chris used his practical experience to validate a communicative approach to teaching and assessment, as explained in his interview:

> 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, [...] So I would say that it was the prevailing culture, and I believed it myself, and I think I 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 “Ahh!” 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.

Such a reliance on communicative principles in assessment yielded two key results. On the one hand, whereas Chris repeatedly commented on language control when providing a running commentary, he seldom took language control into account when he assigned a score to a composition. This reflected his view that mastery of grammar emerged out of the struggle to communicate, which more or less reverses Sam's perspective that control of grammar was a prerequisite to successful communication. On the other hand, perhaps taking the principles of the communicative approach to their logical conclusion, Chris saw grammatical competence as the sign of a very high level of proficiency. From this perspective, his lack of comments on grammar, except, significantly, in essays that he rated relatively high, may reflect his feeling that few of the essays showed a level of proficiency where it was worth commenting on language control, a feeling reflected in his refusal to award a 6 to any of the 60 compositions in the corpus he rated. Thus, dealing with compositions that he viewed generally poorly, Chris focused squarely on the effectiveness with which their authors communicated their ideas.

Chris's teaching principles and experience were useful not only in establishing a general trajectory of language learning, but also in allowing him to identify what students at certain levels of proficiency were capable of. For example, Chris found that creativity in both ideas and language (particularly in the mastery of a range of sentence types) was indicative of high-level writers, and redundancy and topic avoidance were weaknesses of relative beginners. Conversely, teaching experience told Chris what students at a certain level were unable to do, how they could strategically compensate for their shortcomings, and what they needed to work on 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.

Finally, Chris even used his concurrent teaching experience to generate norms, suggesting that he may have been harsh in his judgments during the rating task for the present study because he was teaching an advanced class at the same time and may have unintentionally compared the compositions in the corpus he was asked to rate to the compositions he was regularly receiving from his students.

To the extent that a communicative approach to language teaching is characteristic of much of ESL, as opposed to EFL, teaching, and considering that most ESL teaching takes place in countries where English is the native language of the majority, Chris's use of a communicative approach may be loosely regarded as a cultural trait as much as an outcome of his teaching experiences; he himself refers to the "culture" of the educational institutions he had worked for prior to his participation in the research project as a major formative influence. Otherwise, however, he never ascribed his scoring criteria to his cultural or linguistic background. His experience with placement testing can also be largely subsumed under his teaching experience, because the criteria he used emerged out of a communicative syllabus. At the same time, his familiarity with scoring rubrics in the literature did help him to set native-like proficiency as the standard for a score of 6, and he likewise used his knowledge of the purpose of TOEFL to set such expectations as writers' ability to appropriate a topic to suit their purposes without slipping into irrelevance.

As for Alex, he routinely compared compositions in the corpus he was rating for the present study to external criteria, such as published rating scales and the performance of his students, past and present. At the same time, the internalized scoring rubric Alex brought to the rating task was clearly open to modification, because Alex also spent a good deal of time comparing compositions within the corpus he had been asked to rate:

> I don't believe in criterion referencing in its absolute terms. [...] My personal belief is that in any sort of assessment the norm-referenced concept always comes in at a certain point. I mean if you see somebody meeting certain specific criteria, then there is always the question of how well he has met this particular criteria? OK, I guess this is where the norm comes in. I mean, given two candidates, when both have met a specific criteria, let's say a 5, OK, there is always the question of who has met it more consistently, you know, throughout the whole piece, who has met the criteria, uh, better in a certain aspect, in a certain specific aspect. So you, you are not looking at a criteria at one level, in each criteria you are looking at multiple levels at the same time.

In assembling general scoring criteria for his rubric, Alex was influenced, above all, by his learning and teaching experiences: The curricula in his experience (reinforced by students' performance) stressed the acquisition of word-form and sentence-level mastery at the lower levels, discourse organization and fluency at the next stage, and, eventually, audience considerations, task fulfillment, and the mastery of a range of genres, particularly of the conventions of academic writing. In the interview, he summarized the influence of this developmental scheme on his general scoring criteria as follows:

> Well, as I said, task based is important. I mean, you, it's, it's more like you understand what you are being asked. So interpreting the prompt correctly is a criterion [...] You don't want to train students who don't know what they are supposed to do, but could only display good command of language, I mean that is not what we want. We want them to be able to perform tasks, the more authentic the better. So that is probably, it probably makes sense because, because it discriminates. Language control is probably a more useful factor to discriminate for the weaker students, whereas task completion makes more sense for the, for those who have already crossed the linguistic barrier, but have a good sense of what they are trying to do. Because task is more related to the aim of communication. I mean, why do we want communication? We want to communicate because we want to influence other people. OK? We want to persuade, we want to convince, we want the boss to buy our points. So, I mean, it seems to me that those who have managed to complete the task are usually those who manage the language at a certain level and they can achieve the use of language. Whereas if they fail to do that that's all right, that is something that they can learn later, when they pick up all the bits and pieces, in syntax and morphology and they will come to a better awareness of task.

Given his extensive assessment experience, it is not surprising that Alex also made specific assumptions regarding candidates and test use. For example, having assumed that TOEFL essay topics were prompts designed to elicit samples of writing, he was willing to relax his generally strong emphasis on task fulfillment.

In moving to more specific scoring criteria, Alex, like the other participants, was faced with the task of defining the extremes of his rating scale. Here, he looked for the bottom level not in existing descriptors, but, consistent with his belief in some degree of norm referencing in any assessment situation, in the corpus itself. He found his bottom level during his first pass through the compositions, and used it as the norm for the rest of his rating session. By contrast, he defined the upper end of his scale as the performance of a competent NNS, with the aid of external criteria: the performance of his top students at the university in his native country as his reference group, as well as his perception of the competence of native speakers. Subsequently, Alex refined his scale to separate essays he rated at 3 from essays he rated at 4, both of which fell into his “average” band. If he thought he had seen, among his own university students, the level of writing exhibited by a particular essay, he awarded that essay a 4; if not, he awarded it a 3. This procedure was necessitated by another of Alex’s assumptions arising from previous rating experience, namely, that the sample of 60 compositions was large enough for scores to exhibit a normal distribution. Given this assumption, he expected a large number of compositions to be given a 3 or a 4, and felt that he had to find a way of reliably separating compositions at these two levels. Another consequence of his assumption of a normal distribution of scores was that Alex felt that he could afford to give more 4s than 5s to the essays he was rating:

> 135. [QUOTE] OK. so, is this a 4, my question is this is a 4 or a 5? [QUOTE] There are reasons to mark this one down for trivial errors. But I think uh, it communicates, the piece communicates. There's a badly formed past tense here. [QUOTE] But uh, the errors are consistent and systematic. So, so, uh, yeah, well, I'll give it a 4. *I probably can allow myself to give more 4s than 5s.* So, this is a 4. OK. *[italics mine]*

As shown by the proportional frequencies of his coded comments, Alex was balanced in assessing language and content: In phase III, 9% of his interpretive and 20% of his judgmental strategies had a language focus, and 16% of his interpretive and 15% of his judgmental strategies had a rhetorical-ideational focus. "Reasoning, logic, and topic development," "task fulfillment," "ideas and rhetoric overall," and "language overall" were his principal yardsticks. The first three were usually weighed to provide a first approximation of the score, and language control was usually assessed to refine that score, although it was used more frequently for less proficient essays given Alex's general view of how learners evolved (as discussed above). Alex felt that "below average" essays provided, at best, incomplete answers 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 that showed some control of language 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 fulfillment, global 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 (as demonstrated by his own students at the university level). Finally he judged those essays to be "above average" that amply fulfilled the task, besides being fluent and creative, and gave a 6 only to essays that not only fulfilled all aspects of the task, but were also free of all but minor linguistic errors, and did not rely on formulaic organizational patterns to achieve coherence and cohesion.

Turning to background factors, Alex's assessment experience had little direct influence on the specific scoring criteria he had internalized, although it did teach him to pay attention to such broad traits as language use or organization, and led him to collect crucial contextual information (such as assumptions about test takers and test use) and combine a criterion-referenced approach (embodied in rating scales) with a norm-referenced one (embodied in Alex's repeated comparisons of compositions to each other). It was, instead, the contrasts that he drew between university students and secondary school students, his own experiences as a learner of English, his knowledge of textbook prescriptions, and his familiarity with such theoretical positions as the Teachability Hypothesis that allowed him to specify his expectations.

Relying on his familiarity with both “Western” and “Oriental” cultural traits, constructs that he employed in the interview, Alex could discern different trajectories of language learning, even if his own experience suggested a fairly strong emphasis on grammar. Cultural norms also 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:

> Plagiarism is a cultural thing. For many [ASIAN] 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 an “Oriental” approach to argumentation that, as an examiner, he had frequently observed. By contrast, he 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, which now emerge not merely as products of teaching practice or of ideas about learners’ developmental trajectory, but also as the outcomes 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 SUITABILITY OF A CANDIDATE'S HOMETOWN FOR LOCATING A NEW UNIVERSITY)! You know, this question doesn't exist! So, [THE WRITER] is playing the game, you know what I mean, the game. 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 toward the intercultural aspect of a public exam like this.

Unfortunately, this clash of cultural values begs the question of how Alex decided on accepting or rejecting specific “Western” values, a question that cannot be adequately addressed here. It is important to stress, however, that "cultural background" is individually constructed, meaning that it is impossible to predict from a knowledge of a rater's ethnic origins how he will absorb different influences. This, of course, is as true of writers as of raters, and provides solid justification for keeping the ethnic and liguistic identities of candidates confidential during public examinations.

Finally, 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 term 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.

Jo's scoring procedures resembled Alex's in that they both combined the internalized scoring criteria that they brought to the rating task with criteria developed from reading the corpus of compositions they had been asked to rate. Both took several passes through the corpus of compositions, establishing a rough, 3-point scale first and developing a 6-point scale only after they had read through most of the compositions at least once. However, Jo was even more meticulous than Alex in comparing compositions in the corpus to each other before settling on scoring criteria. As a result, these criteria did not crystallize until the final phase of his rating session, when he was assigning specific scores; moreover, they were clearly derived from the qualities of the corpus of compositions Jo was rating, and were open to modification at any time, as Jo's final reading of essay #63 demonstrates:

> ... Oh, boy. I think [essay #63] is comprehensible. So I think it's a 3-4. ... I'm just going here to take a look at a 3 ... *But even my 3s have had an introduction. ... I look at all the 3s, because so far it looks like my 3s have, at least have the structure of an introduction, body and conclusion. Yeah, look at that, there's another 3. ... Yeah. Topic sentence. Every 1 of my 3s has a topic sentence and some kind of an introduction.* ... But, ... I have to make an exception in this case, I have to give it a 3. Only because it's not, the person is not

> struggling with comprehensibility. I can understand what they are saying. They are struggling with, uh, the language. That's what I said a 3 was. Somebody who is, who can make their ideas comprehensible. It's not difficult to understand what's meant by this at all. It's just that it's not sophisticated. ... But -. Yeah, it's OK as a 3. In comparison to number 77 in particular. So 63 is getting a 3. *[Italics mine]*

The two general criteria that Jo had internalized prior to the rating session were "comprehensibility" and "sophistication," with the former defined as the ability to get a message across and the latter as the ability "to turn a phrase," as well as the ability to take an abstract and detached view and communicate one's ideas in what may be termed an academic style of writing. In applying them, Jo first had to determine the breadth of his scale, combining his knowledge of rating scales with his initial impressions of the corpus of compositions he was currently rating, as shown by comments during the initial phase of his rating session:

> This is number 43 now. [...] OK so here's another one that probably sits somewhere in the low to middle area and of course, the, the thing that immediately strikes me is, "How extensive or how broad is, is my scale?" I know that 1 is the lowest and 6 is the highest and if one wanted to say, you know, 6 could be native speaker and 1 could be a person that can't even communicate. So, again I'm sort of looking to adjust my scale to uh, to this population and thinking OK is 1 the lowest learner in this population? Or is 1 perhaps a learner that, that isn't part of this population?

Eventually, like Chris, who began by pulling short compositions from the corpus to establish his "bottom line," Jo also relied on the corpus to tell him what the attributes of an essay rated at 1 were: He found a good anchor paper in essay #50, and used it to compare other essays with throughout his rating session. On the other hand, he used both his knowledge of existing descriptors as well as an adherence to multiculturalism to define the requirements for his top level, which he did not equate with native-like competence:

> [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.

With the aid of such criteria Jo first established an interim, 3-point scoring scheme for the compositions, and articulated it during the second phase of his rating session, where he started assigning scores 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.

The key criteria, as already discussed, were "comprehensibility," which Jo 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 essays 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.

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 behaviors throughout his protocols. Overall both Jo's interpretive and judgmental comments were weighed heavily toward self-monitoring behaviors, representing 73% of all his interpretive and 52% of all his judgmental strategies, 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 toward a language focus (16% of all coded strategies versus 11% for rhetorical-ideational focus); his judgmental comments were almost evenly split between rhetorical-ideational and language foci.

Although Jo did not articulate the 6-point scale he constructed, he frequently defined his individual scale points, starting with the criteria for getting either the top or the bottom score. He felt that the authors of essays he scored at the lowest level 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 were at least able to present coherent arguments in favor of their stances, even if they could not develop their ideas fully and committed serious linguistic or organizational errors. Although Jo's criteria for getting a score 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 that he thought contained the basic structural elements but were either insufficiently developed or error prone. By contrast, essays that 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 he scored at 4. Finally, essays rated at 5 demonstrated grammatical competence as well as logical organization, and essays rated at the top level brought, in addition, fluency and sophistication to the task. And, as Jo's assessment of essay #128 shows, even if he felt protective toward writers' voices, he certainly expected writers at the top level to produce essays that were essentially error free.

In sum, in scoring the compositions, Jo, like the other participants, relied on broad guidelines, in his case stemming from his ideas of language development, his experiences with teaching and test development, and his assumptions about the purpose of TOEFL. For example, "comprehensibility" originated in Jo's adherence to the principles of communicative language teaching:

> So [emphasis on comprehensibility] 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?

However, in judging the effectiveness with which writers communicated their ideas, Jo relied on his experiences with a process approach to writing instruction, where successive episodes of feedback addressed errors of differing degrees of severity — global errors first, local errors afterward (cf. Burt & Kiparsky, 1972). "Sophistication" as a criterion arose, by contrast, from Jo's experiences with test development, where he was able to observe the performance of writers at various levels in response to specific writing tasks. Jo viewed "sophistication" as crucial to achieving the top score, and defined it as not only the ability to "turn a phrase," but also the ability to take an abstract, or detached, view and to adopt an academic writing style (in contrast with less proficient writers, who were constrained to writing about personal opinions, and in a more experiential style).

Given Jo's understanding of the purpose of the TOEFL essay, the ability to produce sophisticated prose was to be appreciated also because it was a prerequisite for university success. In addition, Jo's stress on the mastery of formal aspects of English writing — although embedded in his conceptualization of comprehensibility and sophistication — reflected his experiences with teaching grammar, as well as his awareness of just how harshly people outside the ESL teaching community could judge a writer's grammatical competence:

> 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.

However, like Alex, Jo also generated specific scoring criteria with reference to the corpus he was rating, through repeated comparisons of essays, because his experience with test development had taught him that it was impossible to predetermine how writers responded to specific prompts. Such a procedure allowed Jo to be flexible without sacrificing consistency, because 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.

Discussion

The design adopted for the study relied on a case-study methodology, with data collected from a wide range of sources used to illuminate how four raters of TOEFL essays with diverse backgrounds rendered specific judgments in a specific scoring situation. Although this approach severely limited the generalizability of the results, the aim of the study was to identify contrasts in the behavior of individuals, and potential sources for such contrasts; the conclusions, therefore, did not need to assume the form of generalizations. Rather, because previous attempts at explaining variability through generalizations came at the expense of reducing both writing assessment and raters' personal and professional backgrounds to single dimensions (comparing, for example, the respective degrees of severity exhibited by NS and NNS raters), the goal here was to identify as many potential dimensions in the process of writing assessment as possible. For the same reason, this study is only a precursor to future studies, on a larger scale, which will have to combine the multidimensional view adopted here with a degree of generalizability that the present design did not allow for. At the same time, given that the participants in the study were selected to maximize variability in their personal and professional backgrounds, similarities exhibited in their behavior could be treated as, at least, hypotheses worth a second look.

Overall, the clearest pattern emerging from the data was that, in the absence of a scoring rubric, the establishment of a link between performance and proficiency was central to the way the four participants generated scoring criteria. Each provided that link through an internalized developmental trajectory for learners of a second language on which they could place the writers of the compositions, and, in constructing specific scoring criteria, they combined their respective trajectories with inferences made regarding the purpose of TOEFL and the attributes of the candidates writing that test. The data collected during the interviews revealed, in addition, that participants constructed their trajectories with reference to their experiences in teaching and, in the case of NNS raters, in learning ESL. Although 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 experiences with rating compositions were helpful principally in the establishment of strategies for reading both individual compositions and a corpus of compositions. Particularly striking was the absence of influence that participants' knowledge of specific rating scales exerted on their scoring criteria, beyond helping to establish the top level in the rating scales constructed by Chris, Alex, and Jo.

All of the participants in the study followed these steps, all agreed that the purpose of 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, in their assumptions of how language could be acquired, and in their definitions of the endpoints of a learning curve.

Such findings have several implications. First of all, given that raters have to process a range of task-specific information, the procedure of establishing rating criteria must be repeated anew for every corpus of compositions that raters assess. Just how fleeting and task-specific rating criteria are is illustrated by Alex's admission, during the interview, that he no longer had any idea why he had given a certain essay a 5, rather than a 4, during the rating session on which he was commenting. Furthermore, although raters in the present study were explicitly instructed not to rely on a scoring rubric, the two most experienced participants, Alex and Jo, 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 by taking additional factors into account. This suggests that rating scales will never be the sole determinants of writing quality in raters' judgments, but will be regarded as only one (even if possibly the weightiest) of several factors that must be taken into consideration. Such a hypothesis could explain why it is that even with a specific scoring rubric interrater reliability can usually be achieved only through extensive discussion of rating criteria and the establishment of anchor papers within a holistic scoring group (as advocated, for example, by Huot, 1996, or White, 1984). It also parallels a growing realization in the literature that (Brown, 1995) "raters have an inbuilt perception of what is acceptable ... formed to some extent by their previous experience." (p. 13; cf. DeRemer, 1998, p. 8; Vaughn, 1991, p. 120)

Such a conclusion also means that two conditions imposed on participants in the present study, the absence of a scoring rubric and the absence of any practical consequences to their decisions, although not representative of authentic assessment situations, should not invalidate the results. The first condition, meant to foreground the influence of prior experience on raters' construction of scoring criteria, essentially eliminated one of the principal sources of information normally consulted by assessors who, according to the findings, consult a wide range of information before rendering decisions. The second condition, necessitated by considerations of fair treatment of test takers, appears to have had a similar effect, in that issues normally grouped under the rubric of consequential validity (the consequences arising out of the specific use[s] that test scores are put to, cf. Messick, 1989) could be (though they were not necessarily) set aside by the participants in the study. Otherwise, based on information extracted from the interviews, only one of the participants altered the procedures he normally followed in response to the experimental nature of the study (by reading compositions only once, instead of repeatedly); the other three treated their task as if it had been an authentic assessment task.

The issue that needs to be addressed in conclusion is the extent to which variability in judgments 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 I had isolated in the initial literature review need to be reassessed. Differences among raters based on ethnicity, culture, and mother tongue — background factors that may be viewed as limiting cases of "extended discourse communities" (Pula & Huot, 1993) — are certainly important, based on the frequency with which participants referred to them in commenting on their own behavior in the interviews. Academic background, by contrast, did not emerge as a significant factor, although whatever unique backgrounds raters have may have been obliterated by their subsequent shared member-ship in the discourse community of ESL teachers and assessors.

Assessment experience (including familiarity with specific scoring rubrics) was likewise limited in its impact, influencing mostly rating strategies, but not the establishment of scoring criteria, beyond directing attention to such general factors as “content” and “organization” in making assessments. In addition, my data, unlike those generated by Pula and Huot (1993, p. 252) in the context of assessing the English compositions of NSs, does not support the view that raters compare the compositions they are rating to a model that they had internalized through their reading experiences. Instead, the most significant sources of variability in the case of three of the four participants in the present study lay in teaching and, in the case of the NNS raters, learning experiences (the fourth participant was guided principally by his work with test development). These were at the root of their 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 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. Finally, although training in assessment procedures can enable a group of raters to render reliable judgments using a particular rating scale, only raters who, largely because of similarities in their teaching experiences, have shared attitudes toward the acquisition of language proficiency — indeed toward the nature of language proficiency itself — are likely to base their judgments on a shared construct of writing proficiency.

References

Alderson, J. C. (1991). Bands and scores. *Review of English Language Teaching* 1(1), 71-86.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, England: Oxford University Press.

Bachman, L. F., & Cohen, A. D. (Eds.) (1998). *Interfaces between SLA and language testing research.* Cambridge, England: Cambridge University Press.

Barnwell, D. (1989). 'Naive' native speakers and judgments of oral proficiency in Spanish. *Language Testing*, *6,* 152-163.

Basham, C., & Kwachka, P. (1991). Reading the world differently: A cross-cultural approach to writing assessment. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 37-49). Norwood, NJ: Ablex.

Belcher, D. (1995). Writing critically across the curriculum. In D. Belcher & G. Brame (Eds.) *Academic writing in a second language: Essays on research and pedagogy* (pp. 135-154). Norwood, NJ: Ablex.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112-140). Cambridge, England: Cambridge University Press.

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

Burt, M. K., & Kiparsky, C. (1972). *The Gooficon: A repair manual for English.* Rowley, MA: Newbury House.

Chalhoub-Deville, M. (1995a) Deriving assessment scales across different tests and rater groups. *Language Testing, 12,* 16-33.

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 l5th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 55-73). Cambridge, England: Cambridge University Press.

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

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

Cumming, A. (1997). The testing of second-language writing. In D. Corson (Series ed.) & C. Clapham (Volume ed.), *Language assessment: Vol.* 7. *Encyclopedia of language and education* (pp. 51-63). Dordrecht, Netherlands: Kluwer.

Cumming, A., Kantor, R., & Powers, D. (2001). *An investigation into raters' decision-making, and development of a preliminary analytic framework for scoring TOEFL essays and TOEFL 2000 prototype writing tasks.* (TOEFL Monograph Series, No. 22). Princeton, NJ: Educational Testing Service.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the writing task. *Assessing Writing, 5,* 7-29.

Educational Testing Service (1989). *Test of Written English guide.* Princeton, NJ: Author.

Erdosy, M. U. (2000). *Exploring the establishment of scoring criteria for writing ability in a second language: The influence of background factors on variability in the decision-making processes of four experienced raters of ESL compositions.* Unpublished master's thesis, Ontario Institute for Studies in Education/University of Toronto, Canada.

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

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning,* 37, 313-326.

Fulcher, G. (1997). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing 14,* 208-238.

Galloway, V. (1977). Perceptions of the communication efforts of American students of Spanish. *Modern Language Journal, 64,* 428-433.

Glaser, B., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge, England: Cambridge University Press.

Guffey, M. E., & Nagle, B. (1997). *Essentials of business communication* (2nd Canadian ed.). Scarborough, Canada: Nelson Canada.

Hadden, B. L. (1991). Teacher and non-teacher perceptions of second-language communication. *Language Learning, 41,* 1-24.

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

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

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 — Writing: Composition, community and assessment* (TOEFL Monograph Series No. 5.). Princeton, NJ: Educational Testing Service.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). The effects of instructions to scorers intended to reduce context effects in essay scoring. *Educational and Psychological Measurement, 43,* 1047-1050.

Huot, B. A. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 237-263.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. A. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 207-236). Cresskill, NJ: Hampton Press.

Huot, B. A. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*(4), 549-566.

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

Janopoulos, M. (1993). Comprehension, communicative competence, and construct validity: holistic scoring from an ESL perspective. In M. A. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 303-325). Cresskill, NJ: Hampton Press.

Kaplan, R. B. (1966). Cultural thought patterns in intercultural education. *Language Learning, 16,* 1-20.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers’ background. *Language Learning, 46,* 397-437.

Kroll, B. (1998). Assessing writing. In W. Grabe (Ed.), *Annual Review of Applied Linguistics, 18,* 219-240. New York: Cambridge University Press.

Kroll, B., & Reid, J. (1994). Guidelines for writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing, 3,* 231-255.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12,* 54-71.

McNamara, T. F. (1996). *Measuring second language performance.* New York: Longman.

Messick, S. (1989). Validity. In Linn, R. L. (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Milanovic, M., Saville, N., & Shuhong, S., (1995). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 92-111). Cambridge, England: Cambridge University Press.

Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis: A source book of new methods.* Newbury Park, CA: Sage.

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education, 3*(3), 285-296.

Pienemann, M. (1986). Psychological constraints on the teachability of languages. In C. W. Pfaff (Ed.), *First and second language acquisition processes* (pp. 103-116). Rowley, MA: Newbury House.

Picnemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition based procedure for second language assessment. *Studies in Second Language Acquisition, 10(2),* 221-243.

Pollitt, A., & Murray, N. L. (1995). What raters really pay attention to? In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 74-91). Cambridge, England: Cambridge University Press.

Prior, P. (1995). Redefining the task: An ethnographic examination of writing and response in graduate seminars. In D. Belcher & G. Braine (Eds.) *Academic writing in a second language: Essays on research and pedagogy* (pp. 47-82). Norwood, NJ: Ablex.

Pula, J. J., & Huot, B. A., (1993). A model of background influences on holistic raters. In M. A. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 237-265). Cresskill, NJ: Hampton Press.

Purves, A. (1984). In search of an internationally valid scheme for scoring compositions. *College Composition and Communication, 35*(4), 426-438.

Raimes, A. (1998). Teaching writing. In Grabe, W. (Ed.), *Annual Review of Applied Linguistics, 18,* 142-167. Cambridge, England: Cambridge University Press.

Santos, T. (1988). Professors' reactions to the writing of nonnative-speaking students. *TESOL Quarterly, 22*(1), 69-90.

Swales, J. (1990). *Genre analysis: English in academic and research settings.* Cambridge, England: Cambridge University Press. 1990.

Sweedler-Brown, C. O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing, 2,* 3-17.

Tedick, D., & Mathison, M. (1995). Holistic scoring in ESL writing assessment: What does an analysis of rhetorical features reveal? In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 205-230). Norwood, NJ: Ablex.

Trimble, L. (1985). *English for science and technology: a discourse approach.* Cambridge, England: Cambridge University Press. 1985.

Tyndall, B., & Mann-Kenyon, D. (1996). Validation of a new holistic rating scale using Rasch Multi-faceted Analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39-57). Norwood, NJ: Ablex.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49,* 3-12.

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

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

White, E. M. (1984). *Teaching and assessing writing* (2nd ed.). San Francisco: Jossey-Bass.

Williamson, M. M. (1988). A model for investigating the functions of written language in different disciplines. In D. A. Joliffe (Ed.), *Advances in writing research. Volume* 2: *Writing in the academic disciplines* (pp. 89-132). Norwood, NJ: Ablex.

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL
programs and services, use one of the following:

Phone: 609-771-7100
Email: toefl@ets.org
Web site: www.toefl.org

73313-025347 • Y44E.05 • Printed in U.S.A.
I.N. 995738