UNIVERSITY OF MURCIA

International Journal
of
English Studies

IJES

www.um.es/engphil/ijes

The Influence of Prior Experience on the Construction of Scoring Criteria for ESL Compositions: A Case Study

M. USMAN ERDOSY[*]
Ontario Institute for Studies in Education of the University of Toronto

ABSTRACT

Before a principled explanation of variability in raters' judgements of ESL compositions can be offered, the process of constructing scoring criteria and the manner in which prior experience enters this process must be analyzed. Therefore, utilizing protocol and interview data collected in the context of a comparative study, a case study will describe how one experienced rater dealt with the following operations while assessing a corpus of 60 TOEFL essays: establishing the purpose of assessment, developing a reading strategy to deal with a corpus of essays, and collecting context-specific information. Within each operation, the influence of background variables such as teaching and assessment experience will be examined, particularly on determining what type of information to collect, and on articulating expectations concerning test takers, test scores and the textual qualities of essays. The results of the study will be used to specify directions for future research into explaining inter-rater variability.

KEYWORDS: background influences, case study method, ESL writing, rater variability, think-aloud protocols, writing assessment.

[*] ***Address for correspondence:*** M. Usman Erdosy, Ontario Institute for Studies in Education of the University of Toronto, Modern Language Centre, 252 Bloor Street West, Toronto, Ontario, Canada M5S 1V6, e-mail: uerdosy@oise.utoronto.ca

IJES, vol. I (2), 2001, pp. 175-196

I. INTRODUCTION

Variability in raters' judgements ofcompositions in English as a Second Language (ESL), raises questions conceming the validity of performance-based writing assessment, if it is interpreted (following Milanovic, Saville, & Shuhong, 1995: 93) as reflecting the absence of a uniform construct as the object of measurement. The fact that variability exists even when raters are instructed to use rating-scale descnptors shows, in Vaughan's (1991: 120) words, that raters "do not, like computers, intemalize a predetermined grid that they apply uniformly to every essay." Such behaviour may be inherent in the assumption that raters of compositions are "readers,"(cf. Huot, 1990; Janopoulos, 1993; Kroll, 1998; Purves, 1984), bringing prior experience to rating tasks. However, this view obliges researchers to explore the process raters follow in constructing scoring criteria for ESL compositions, and identify the manner in which raters' prior experiences enter this process, thus contributing to a principled explanation of inter-rater variability. These will be the objectives of the present study.

Unfortunately, recent studies have been concemed mostly with the outcomes, rather than with the process, of ESL writing assessment: identifying textual characteristics that raters focus on, and/or measuring the level of severity reflected by the scores raters assign. These studies have, in addition, treated prior experience one-dimensionally, considering such background variables as mother tongue, academic orientation, level of assessment experience, age, or gender in isolation. As a result, while several background variables have been identified in the literature as potential influences on scoring criteria, the conclusions offered have been not only limited but frequently inconsistent. Thus, for example, both Santos (1988) and Vann, Lorenz, and Meyer (1991) asked faculty members in different university departments to rate doctored ESL essays, and attempted to relate raters' level of severity to their prior experience. While their results agreed in that social science faculty in both studies were more lenient than natural science faculty, Santos (1988) found age and mother tongue to be significant additional factors in establishing a rater's level of severity, while Vann et al. (1991) pointed, instead, to gender. However, without considering how prior experience, such as academic orientation, actually translates into specific expectations and how, and at what point, such expectations play a role in the rating process, the findings conceming variability, such as those just cited, will be impossible to explain, and contradictions between them impossible to resolve.

At the same time, previous studies have laid the groundwork for an in-depth analysis of the process of ESL writing assessment. These include the identification of critical variables in the process of assessment (Hamp-Lyons, 1990; Kroll, 1998); the development of techniques of data collection and data analysis, particularly the elicitation of concurrent verbal protocols and their coding (Cumming, 1990; Pula & Huot, 1993; Vaughan, 1991); and typologies of decision-making behaviours (Cumming, 1990), decision-making sequences (Milanovic et al., 1995), and textual features raters attended to (Cumming, 1990; Huot, 1993; Vaughan, 1991). In addition, a study by Pula and Huot (1993; Huot, 1993) provided a detailed treatment of the role of prior experience, including personal background, professional training, and work experience, in the assessment of English, although not ESL, compositions. Its key conclusions were that raters in the study relied, above all, on their reading experiences to form idealized images of "good writing", that "content" and "organization" were their key criteria in determining what "good writing" was, and that they assessed English compositions by comparing them to their ideals of "good writing" (cf. Gorman, Purves & Takala, 1988). However, differences between L1 and L2 writing (Silva, 1993) and L1 and L2 writing assessment (Hamp-Lyons, 1991a) suggest that these conclusions regarding the processes raters follow in writing assessment are not easily transferable to the context of assessing ESL compositons, especially since Pula and Huot's study ignored background variables, such as mother tongue and cultural background, which were not relevant to the assessment of L1 writing, but are highly relevant to the assessment of L2 writing (cf. Li, 1996).

II. OBJECTIVES OF THE PRESENT STUDY

Given the focus of previous studies on the outcomes of assessment, and their generally one-dimensional view of prior experience (with significant exceptions noted above), the objective of the study reported later was to lay the foundations for a principled explanation of variability in raters' judgements of ESL compositions. The aim, specifically, was to identify key operations within the assessment process, to specify relevant background variables in raters' prior experience, and to identify both the instances when prior experience influenced the assessment process and the manner in which it did so.

In its first stage (Erdosy, 2000), the study involved analyzing the behaviour of four raters, in order to identify contrasts in the manner in which they approached a single rating task (assessing 60 TOEFL essays) and to explore the influence of prior experience on the observed contrasts. Three key operations were highlighted in the assessment process as particularly influenced by prior experience: establishing the purpose of assessment, developing reading strategies, and collecting information in order to generate scoring criteria specific to a particular rating task. Within prior experience, in turn, personal background, professional training, and work experience (cf. Pula & Huot, 1993), as well as mother tongue and cultural background, could be identified as critical background variables.

Once key contrasts between participant raters were identified, both long-term and short-term options emerged for follow-up studies. The ultimate objective, naturally, was a principled explanation of variability in the judgements of raters of ESL compositions, using the contrasts tentatively identified on the basis of a limited comparative study, and involving a larger sample of raters. The short-term option, adopted here, was to construct a case study detailing the assessment process followed by one of the raters involved in the study, as well as the influence of prior experience on that process. Such a study would be descriptive, and would not directly address the question of inter-rater variability. However, by demonstrating the complexities of the assessment process, it could identify fruitful directions for a study of inter-rater variability.

Consequently, drawing on the data collected, and on the contrasts identified, during the comparative study (Erdosy, 2000, in turn taking data from concurrent verbal protocols furnished for a larger study by Cumming, Kantor, & Powers, in press), my objective here is to present a detailed description of the assessment process followed by one experienced rater of ESL compositions, guided by the following research questions:

A. *How did the participant rater conceptualize the purpose of performance-based writing assessment and what role did background variables play in this operation?*

B. *What reading strategies did the participant rater establish to deal with both individual compositions and the corpus ofcompositions he was asked to assess, and what role did background variables play in this operation? In particular,*

– How many times did the participant rater read individual compositions in a corpus?

– In whatprincipled order did theparticipant rater read compositions in a corpus?

C. *What information did the participant rater seek when generating specific scoring criteria and what role did background variables play in this operation?*

III. RESEARCH DESIGN

III.1 The Participant

Alex was an East Asian doctoral student in second language education at a North American university, in his late 40s at the time of the study. He was invited to participate because of the extent of his experience with both teaching and assessing ESL writing in his native country. His 12 years' teaching experience spanned the secondary and tertiary levels ofeducation, in addition to teaching English for Special Purposes (ESP). Besides frequently conducting classroom evaluation, he had conducted placement testing at the university level, had served as an assessor for a nation-wide English examination authority, was familiar with a wide range of scoring rubrics for ESL writing assessment (referring explicitly to rubrics published in Hamp-Lyons, 1991b, and Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey, 1981), and had been involved in rater training. Further, as a non-native speaker of English, Alex had experienced assessment from the perspectives of both assessor and test-taker, and was an experienced writer in both his first and second language.

111.2 Data Collection

The data conceming Alex was collected frorn four sources. The principal source of information for Alex's ratings consisted of concurrent verbal protocols Alex fumished while assessing a corpus of 60 TOEFL essays in the context of an ETS-funded study into raters' decision-rnaking (Curnrning, Kantor & Powers, in press). The essays had been written during four (then) recent adrninistrations of the TOEFL at a North Arnerican site, with 30 rninutes allotted for task cornpletion. They ranged in length frorn one typewritten line to one and a half (single-spaced) typewritten pages; however, the topics they responded to (numbering four) cannot be identified due to a confidentiality agreernent goveming the use of TOEFL data for the study.

Alex was not informed of either the scores that had been originally assigned to the essays, or the identity of the authors of the cornpositions. Instead, he was instructed to assess the corpus anew, using a 6-point scale. In doing so, he was invited to refer to rating scales he was familiar with, if he felt that these had influenced his criteria; conversely, he was requested not to base his assessrnents on any of these scales, but to construct his own scoring criteria. This experimental condition was irnposed to focus the study on exarnining the influence of prior experience on Alex's rating process, rather than on validating an existing rating scale. The other result of the experimental nature of the study was that Alex's ratings had no practical consequences. However, while this rnay have influenced his assessrnents, the fact that he repeatedly referred in a follow-up interview (see below) to his understanding of the nature ofthe TOEFL as one of the key factors goveming his construction of scoring criteria suggests that the influence of the experimental nature of the assessrnents was minirnal; an additional reason for this assurnption is that Alex repeatedly asserted in a follow-up interview that the procedures he followed were those he would have ernployed in authentic assessrnent situations.

In fumishing the concurrent verbal protocols, Alex was instructed to cornrnent aloud on the cornpositions, speaking continuously, speaking in English "as rnuch as he could" (Cumming, Kantor & Powers, in press: 78), and avoiding speech filters such as "uh." Apart frorn a request to report his first irnpressions and how they rnay have influenced his ratings, he was instructed to follow whatever procedures came naturally to hirn. There was no time lirnit set for the task, which Alex executed at horne, in a single, 3-hour session, following a 45-rninute practice session in the use of concurrent verbal protocols using a simple cognitive task. The concurrent verbal protocols were taped and subsequently transcribed. They were then coded by the researcher, using a scherne originally developed by Curnrning (1990; cf. Curnming, Kantor & Powers, in press), and the coded transcript, along with the scores Alex assigned to the cornpositions, provided the first source of information.

The second source of information was represented by Alex's answers to a questionnaire (cf. Appendix A), which elicited information on his personal background, professional training, work experience, and reflections on the scoring task itself. It also asked Alex to identify the three most irnportant factors in his past experience that he felt rnay have influenced his assessments, and the three qualities that "make for especially effective writing in the context of a compositions examination". As requested, Alex answered the questionnaire immediately after completing the rating task, and the information contained therein acted as a check on statements that Alex made in the course of the interview (see below).

The third source of information consisted of an interview, which served a dual purpose. It invited Alex to comment on his behaviour during the rating session in the light of his prior experience, as well as to assess the accuracy of my analyses of this behaviour. It consisted, initially, of a prompted recall protocol which asked Alex to comment on transcripts of protocols concerning 12 of the 60 compositions he had rated, and which took 90 minutes to complete. The only instmction asked Alex explicitly to comment on the protocols in light of his background "as a learner, teacher and/or assessor of English, ESL, or any other language".The second part of the interview adopted a semi-stmctured format, with the stmcture provided by the analysis of Alex's concurrent verbal protocols and his statements in the questionnaire. It was designed to elicit information on the key steps of the rating process and to explore the background variables that a review of the literature had suggested as critical to understanding the criteria that experienced raters of ESL compositions constructed and/or applied in the rating process. The second part of the interview also took 90 minutes to conduct. As was the case with the concurrent verbal protocols, the interview, which took place in the researcher's office, was taped and transcribed by the researcher.

Then, at the conclusion of the study, Alex was furnished with a draft of the analysis of his behaviour, and requested to assess the degree to which he felt that the interpretation of his behaviour was accurate. Such a member-check, along with the use of multiple sources to facilitate triangulation, acted as quality control on the analyses conducted.

IV. FINDINGS

IV.l. Research Question A: How did Alex Conceptualize the Purpose of Performance-based Writing Assessment and What Role did Background Variables Play in this Operation?

In answering the questionnaire Alex explicitly referred to "knowledge and understanding of the relationship between proficiency and performance" as a key influence on his assessment of compositions. Later, in his interview, he clarified that in assessing a composition, he was essentially seeking to infer proficiency from performance:

> 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. ... 1 have, actually taught graduation level for many years, and 1 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 operationalizing "proficiency", Alex repeatedly associated it with language control at the sentence level and text organization; to these he added task fulfilment as a secondary criterion. Then, defining "performance" in this instance as an essay written in response to a TOEFL prompt, under the usual conditions specified for the TOEFL, Alex identified the purpose of his assessments in the present rating situation as providing a score weighing both proficiency (as understood above) and, to a lesser extent given the nature of the test, task fulfilment:

> ... my question was "Whether or how much 1 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 1 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, 1 felt, at one point, not very important. So, I would still try to give some ofthese 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 providing a case in point]

Alex also commented that in performance assessment "you cannot go beyond what the performance suggests". However, in light of his attempts to, in his words, "associate a certain performance with a certain level ofproficiency", this statement must be taken to indicate Alex's attitude to considering the impact of situational factors, such as time pressure or topic effects, on performance. Based on his protocols, Alex was clearly aware of the effects of situational factors, noting, for example, that the writers of some essays appeared to be writing under time pressure. Yet, as shown by his comments conceming a clearly unfinished essay (# 104), he took compositions at face value:

> 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.

All in all, taking a performance at face value did not, for Alex, preclude inferring language control from indicators in an isolated performance, but it did preclude speculating on what a writer's level of performance may have under conditions more favourable than he/she was exposed to in any particular assessment situation.

Regarding the influence of background variables on Alex's definition of the purpose of his assessments as judging language control, text organization, and task fulfilment, his assessment experience created, initially, an awareness of the purpose of the assessment instrument that he was dealing with. He explained that his policy of giving language control and text organization greater weight than task fulfilment stemmed from his perception ofthe TOEFL as a test whose international nature necessitated that task requirements be framed in very general terms:

> I understand that the contemporary writing researchers seek to avoid asking display questions, but Idon't see how this is possible [in] a public exam like TOEFL, frankly, I don't see how it is possible. It is much more possible if we can contextualize ...the prompt, to, you know, individual level. But then the question ofcomparison comes, you know, like reliability. So, it is not aquestion that, I think, can be easily tackled in a public exam as large as TOEFL.

The same conceptualization of the TOEFL as inviting a display of language control also enabled Alex to downplay the impact of essay topics on a test-taker's performance, a policy consistent with his policy of taking performances at face value.

Additionally, the fact that Alex understood the uniqueness of every rating task, requiring the construction of situation-specific scoring criteria, itself came from his experience as an assessor of second language writing:

> In all the exams that I marked, including institutional exams, you know, in department, in school, we all used a marking scheme and in fact in later years, at the university level, we decided that a, uh, universal marking scheme doesn't work any more. It was more like you use a marking scheme every time for a specific prompt or task. You devise a new marking scheme every time and you don't use the same one. We have this problem, because when you deal with different document types you realize that you do need different..., uh, it's better, I mean it's not like that it [the other] won't work, but it's better, it's more reliable if you devise a specific scheme for that particular task.

It is this understanding of the situation-specific nature of assessment that guided Alex to collect specific types of information, such as the nature and purpose of the assessment instrument he was administering, and determining what types of information to collect represented one of the means for prior experience to enter the assessment process.

The second means of entry was offered in the present rating situation by Alex's need to operationalize key concepts, and this is exemplified by the influence of teaching experience on Alex's assessments. For example, Alex recalled how teaching in a task-based curriculum at the university level in his native country helped him to define task fulfilment not only as answering the questions posed by essay topics, but also as developing audience awareness and cultivating an academic tone:

> Well academic task, I mean, at least you have a role to assume, you have to know who you write to. you have to know why you are writing it and you try to realize these in the writing... they are also expected of secondary school students but that expectation is usually not very realistic in the sense that you more or less bother with syntax and word choice and word formation problems more than, you know, audience analysis or audience orientations and genre.

Alex likewise used his teaching experience to isolate aspects of language control that could be used to measure a writer's level of "proficiency": two such criteria were the level of flexibility in paragraph structure and the variety of cohesive devices used, with low-level writers showing little or no control, mediocre writers relying on a limited number of formulaic devices, and advanced writers showing flexibility and variety. Such criteria were based on Alex's exposure to students at both the secondary and the tertiary level which afforded him first-hand experience of how learners progressed, and how their performances in classroom tests related to their proficiency level.

On a more fundamental level, Alex not only operationalized scoring criteria, but also scoring procedures, since his practice of inferring language control from isolated aspects in a performance assumes the existence of a developmental trajectory for second language learners, a trajectory that was suggested to him by a convergence of teaching experience with theoretical principles. He referred, in particular, to the influence on his thinking of Pienemann's (1986, 1998) Teachability Hypothesis:

> 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 profíciency 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.

IV.2. Research Question B: What Reading Strategies did Alex Establish to Deal with Both Individual Compositions and the Corpus of Compositions He was Asked to Assess, and What Role did Background Variables Play in this Operation?

The reading strategies Alex established not only influenced the scores he assigned, but also provided the context in which statements concerning the compositions in his concurrent verbal protocols had to be interpreted. For this reason, their analysis formed a key component of the present study.

Using the taxonomy adopted by Milanovic et al. (1995), Alex's overall strategy could be roughly classified as a "principled two-scan read". Following the initial scanning of a few compositions for length and appearance, Alex read the entire corpus without altering the order in which he had found it. Then, having assigned tentative scores and established a rudimentary rating scale, he reread the compositions, this time grouped by the scores he had assigned. He continued reading the compositions, and comparing them both within and across groups, until he was satisfied that the groups he had established were internally consistent and clearly distinguishable from one another, at which point he finalized his scores. Readings of compositions during the final stage could be terminated as soon as Alex was satisfied with the score he assigned, suggesting that he operated in a hypothesis-testing mode, an interpretation he concurred with during the interview. An example of Alex's reading strategies is fumished by the following protocol, conceming essay #144:

> **(initiai reading)** #144. [QUOTE] What's this? [QUOTE] Oh, my goodness! The level of information, that is, the structure; suddenly, I think this is a 1+.
>
> **(second reading)** Now, #144 is a 1 and I have to go back to all the 1's. OK, #68 [QUOTE] So, that is definitely a 1... so, then #144 is ... probably a 1, too. But at least he's answering the question, although it's short. So, that may be a 2. So, go back to it later.
>
> **(third reading)** #144. [QUOTE] It's [GOT] some stupid spelling errors ... but it's not like the other [essays Alex rated 1] ... so, I'm gonna upgrade this to a 2.

Although Alex's strategy goes counter to the approach of reading compositions once, and reading them rapidly, which has been recommended for holistic scoring (cf. Hamp-Lyons, 1991c: 243; Vaughn, 1991: 113), Alex felt that it was firmly grounded in his prior experience with assessment. In particular, in the assessment sessions Alex had participated in raters were instructed to factor the range of proficiency displayed by a corpus into scores for individual compositions, on the assumption that the scores had to be normally distributed. The impact of such an assumption may be seen in the following protocol, showing Alex was not averse to downgrading essays to achieve a normal distribution of scores:

> 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 comrnunicates. 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 4's than *5's*. So, this is a 4. OK. [*italics* mine]

Obviously, arequirement to produce normally distributed scores clearly dictates multiple readings and the sorting of cornpositions into piles as a way of ensuring consistency, a habit reinforced by Alex's own dissatisfaction with purely criterion-referenced assessment:

> My personal belief is that in any sort of assessment the norm-referenced concept always comes in at a certain point. I mean ... ifyou see somebody meeting certain specific criteria, then there is always the question ofhow 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 ofwho 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.

Once again, it is possible to see the operation of prior experience through framing expectations (such as a normal distribution of scores), and through directing the collection of specific information. However, in the present study an additional factor may have been the nature of the assessment task Alex was undertaking. As Hamp-Lyons (1991c: 244) mentions, one weakness of rapid, holistic reading is that raters are usually unable to rationalize their scores, yet, in this case, Alex was asked to provide precisely such rationalizations, and, in addition, had to operate in the absence of a scoring rubric. If this is true, then Alex was once again influenced by the perceived purpose of assessment; although still recognizing that he was to simulate a rating session involving TOEFL essays, he now acknowledged that his assessments were for experimental purposes.

IV.3. Research Question C: What Infornation did Alex Seek When Generating Specific Scoring Criteria and What Role did Background Variables Play in this Operation?

Having determined the purpose of his assessments for the present study, developed such expectations as a normal distribution of scores in a corpus of 60 compositions (an expectation he later realized may have ben unrealistic), and established a reading strategy, Alex proceeded to make assumptions conceming test takers. To wit, based on his knowledge of the TOEFL, he inferred that the test takers were applying for admission to North American universities. The expectation that arose from this assumption was that test takers would have to become familiar with American cultural realities sooner or later. Consequently, Alex did not feel the need to ascertain the impact of culturally biased essay topics. He could even refer to time-honoured traditions in examinations for public offices in his native country, thus bringing his own cultural background into play, in deliberately ignoring such information even where it may have been made available by the test takers themselves:

> As a reader, when you read something you do want to seek contact with the writer, sometimes even in terms of personal contact. But, in assessment we try not to do that so that we won't be biased against certain types of candidates. The origin of exams, especially in [MY COUNTRY] was to decontextualize candidates, so that their talents could be assessed in terms of the talent, you know, their writing ability, their eloquence. ... Not where they come from, whether they come from a poor village, or they are a farmer's daughter, uh, son. I mean, that's why you have the exam, right? This is testing. I think it's a very revealing remark, you know, like "we assess because we want to decontextualize other factors." So, what's the point oftesting [OTHERWISE]?So, uh, I do hold a more detached view.

On a different level, Alex also observed that the better pupils among those he had taught at the university level could deal with awkward questions in examinations through assuming a persona and stances that they may not necessarily have believed in. Consequently, he found additional justification for ignoring test takers' ethnic and cultural backgrounds. In another example of framing expectations based on prior experience, the performances exhibited by first and second-year university students in Alex's native country served as key benchmarks: essays in the corpus he assessed for the present study were assigned scores of 3 or lower if they didn't match the performances he had observed among his own students, and scores of 4 or higher if they did.

Beyond generating such expectations, however, assessment experience did not influence the construction of specific scoring criteria. Instead, Alex began with the instruction, given to him at the outset of the experimental rating session, to use a 6-point scale. Then, he relied on his teaching experience to establish a rudimentary developmental trajectory for leamers of ESL. This began with the mastery of basic structures at the sentence level, continued with the mastery of discourse competence and the acquisition of a degree of fluency, followed by a gradual abandonment of formulaic organizational pattems in the achievement of overall coherence and cohesion, and culminating in the mastery of a range of genres, and the ability for extended argumentation along with the elimination of most errors. Alex also recognized, based on his teaching experience, that language control was a useful yardstick particularly for lower levels ofproficiency (cf. Pollitt & Murray, 1995); thus, it is not surprising to see an inverse relationship in his protocols between the frequency of comments in his protocols conceming language control and the scores he assigned:

> 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 a t a certain level and they can achieve the use of language.

In the final step, Alex sought to define his scoring criteria more specifically, and here he once again relied heavily on teaching experience, although at this point his procedures became more haphazard, with some of his scoring criteria clearly articulated and others nebulous. One of the more clearly articulated criteria involves Alex's requirements for awarding a score of six; teaching experience, combined with his perception of native speakers' competence, suggested that even at the highest level, language errors were bound to occur:

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

Examples of criteria in need of some refinement include Alex's attitude to originality and plagiarisrn, which have been shaped by Alex's cultural background:

> 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?

Yet, to underscore the dangers of generalizing frorn a rater's ethno-cultural background, Alex was just as ready to accept what he termed "Western" cultural values such as a dislike of sweeping, unsupported generalizations:

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

All in all, the scoring critena that emerged out of such influences relied on an initial tripartite division: Alex awarded "below average" compositions a 1 or a 2, "average" ones a 3 or a 4, "above average" ones a 5 or a *6*. Overall, Alex felt that "below average" cornpositions provided, at best, a minimal response; among these, compositions deficient in language control and organization were awarded a 1, those showing sorne control of language and/or some awareness of basic structural requirements received a 2. "Average" compositions provided either incomplete or one-sided argurnents, with those displaying limited task fulfilment and global errors getting a 3, while those fulfilling the task and free from global errors, but suffering from logical flaws, irrelevance, or inadequate development got a 4. "Above average" compositions amply fulfilled the task, besides being fluent and creative, with a *6* awarded to essays which were free of all but minor linguistic errors. One additional distinction that Alex made was between compositions that matched the performance of his students at university-level writing classes in his native country, which got a 4, and cornpositions that fell just short of this level, which got a 3. Expecting a normal distribution of scores, and thus the majority of compositions to be awarded a 3 or a 4, Alex felt the need to make such a distinction in cases where his usual cnteria failed him.

V. DISCUSSION AND IMPLICATIONS

In assessing a corpus of compositions without relying on a scoring rubric, Alex performed the following key operations: identifying the purpose of assessment; developing a reading strategy; collecting context-specific information (including instructions given to raters) concerning test takers, test use and test administration; and generating specific scoring criteria. It is in executing these operations that prior experience influenced Alex, suggesting that manipulating the context for these operations and examining raters' reactions to changing conditions may be how future, experimental studies concemed with variability could produce the most furitful results. Changes in raters' decisions regarding the information to be collected, and in their expectations conceming test takers, test scores and the textual qualities of essays, which, in tum, could be translated into scoring criteria would be particularly important to specify.

Another consequence ofthe need to collect extensive context-specific information during a rating session may be that even if raters are instructed to rely on a specific scoring rubric, that rubric will represent only one (although, ideally the most important) piece of information that raters will heed in the assessment process. This should explain the finding, reported at the outset, that raters (whether of speaking proficiency – cf. Brown, 1995 – or of writing proficiency – cf. Vaughan, 1991) do not mechanically apply a scoring rubric even if they are instructed to use one. This, in addition, does not even begin to take into account the problem that rating scale descriptors, in Alex's experience, always underspecified the criteria associated with any given point on a rating scale:

> I think a descriptor is not helping much, in my view [BASEDON] past experience with descriptors for holistic rating. Let me put it this way. The problem with a descriptor is the assignment of proportions of different aspects of a piece of writing. I mean they always cross, interact themselves. You know, you have to look at the interactional effects between the various aspects that eventually you come down with a simple number. So that while the descriptors are there as a guideline, more like, and they have no substantial help in terms of deciding whether [a paper] is a 4, because a 4 is sometimes short in grammar but strong in ideas and then you have organization and so forth. So what is a 4? A 4 can mean a host of longs and shorts of many aspects.

Finally, in light of the need to gather context-specific information, the process of establishing scoring criteria would have to be repeated anew for every rating session. Indeed, this is a conclusion that Alex himself reached during his time as an assessor (cf. p. 11, above). It is this factor that may explain within-rater variability: unless the specific circumstances of one rating session can be faithfully replicated in another, there is no reason why a rater – absorbing different sets of contextualized information in the two rating sessions – should give an identical rating to the same composition in different contexts. A dramatic confirmation of this carne in the following statement made by Alex in the context of his prompted recall protocols during the first phase of his interview:

> **... now that I am looking at it, it would still be a mystery to myself why I gave a particular one a 5 ora 4. You see when I am looking at it "That's a second reason... makes good sense." So I gave this [composition] a 5. It's difficult to explain now why I gave this one a 5.**

All this is not to imply that rater reliability is unattainable in performance-based assessment (assuming, for the moment, that such a suppression of divergent opinions, is indeed desirable). Studies, such as those conducted by Cumming (1990) and Weigle (1994), show that rater training can significantly reduce variability in raters' judgements. If one takes the process one step further and allows raters to negotiate their rating criteria (as suggested by Huot, 1996 and White, 1984, and also by Alex's description of the frequently heated debates between raters in the standardized assessment program he was involved in), rater reliability could be improved further still. However, the scope for the local negotiation of scoring criteria is greatly reduced for a standardized test like the TOEFL, since consistency at the level of a holistic scoring task group (to use Pula and Huot's, 1993, term) would come at the expense of comparibility across groups. An altemative solution suggested by this research would be to expand rater training beyond the use of rating scales and beyond the use of anchor papers to a systematic consideration of the entire range of factors identified here as bearing on the establishment of scoring criteria – the characteristics of test takers, the purpose of a test, and the baggage of intemalized scoring criteria that every rater carries to a rating task.

Acknowledgement

I would like to thank the reviewers for their comments which greatly improved the focus and quality of my paper.

REFERENCES

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

Cumming, A. (1990). Expertise in evaluating second language composition. ***Language Testing,*** 7 (1), 31-51.

Cumming, A., Kantor, A., & Powers, D. (In press). ***An investigation into raters' decision-making, and development of a preliminary analytic framework for scoring TOEFL essays and TOEFL 2000 prototype writing tasks.*** (To appear in the TOEFL Monograph Series). Princeton, NJ: Educational Testing Service.

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-rnakingprocesses** *offour experienced raters ofESL compositions.* M.A. Thesis. Toronto, ON: Ontario lnstitute for Studies in Education/University of Toronto.

Gorman, T. P., Purves, A. P., & Takala, S. (1988). The development of the scoring scheme and scales. In T. P. Gorman, A. P. Purves, & R. E. Degenhart (Eds.), *The IEA study of written composition 1: The international writing tasks and scoring scales* (pp. 41- 58). Oxford: Pergamon Press.

Hamp-Lyons, L. (1 990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991a). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5-15). Norwood, NJ: Ablex.

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

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

Huot, B. (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: Hampton Press.

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

Li, X-M. (1996). *"Good writing" in cross-cultural contexts.* Albany, NY: SUNY Press.

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

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

Pienemann, M. (1998). *Language processing and second language development: Processability theory.* Amsterdam: John Benjamins.

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

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 intemationally valid scheme for scoring compositions. *College Composition and Communication,* 35 (4), 426-438.

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

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarferly,* 27 (4), 657-675.

Vann, R., Lorenz, F., & Meyer, D. (1991). Error gravity: Faculty response to errors in the written discourse of non-native speakers of English. ln L. Hamp-Lyons (Ed.), *Assessingsecond language writing in academic contexts* (pp. 181-195). Norwood, NJ: Ablex.

Vaughan, C. (1991). Holistic assessment: what goes on in the raters' minds? ln 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 wrifing* (2nd ed.). San Francisco: Jossey-Bass.

Appendix A:

Questionnaire

The purpose of this questionnaire is to gather background information which will be related to the data you will generate in the think-aloud protocols while you assess the ESL compositions for this research. Please note that the aim of the research is not to evaluate your performance, but, rather, to understand it more fully. As with other data generated by the project, your identity will remain confidential.

I Your Assessments

1. What are the three most important factors influencing your assessment of second language compositions?

i) ____

____

____

____

ii) ____

____

____

____

iii) ____

____

____

____

To what extent do any assessment scheme(s) (e.g. rating scales, checklists, etc.) influence you in assessing compositions? Please circle the number that best corresponds to your answer.

1	2	3	4	5
not at all		slightly		a great deal

If you indicated any degree of influence (i.e., circled 2, 3, 4 or 5), please elaborate on the extent and nature of that influence.

____

____

____

____

4. What three qualities do you believe make for especially effective writing in the context of a composition examination?

i) ____

____

____

____

ii) ____

____

____

____

iii) ____

____

____

____

II personal **Profile**

5. Gender: Male ____ Female ____

6. Age: ≤ 30 ____ 31-40 ____ 41-50 ____ > 50 ____

III Current Professional Status

7. Your **current role(s)** Assessor ____ Teacher ____ **Administrator** ____

Student ____ Other **(specify)** ________________

8. The **context(s)** **English** ____ ESL ____ EFL ___

ESP ____ Other **(specify)** ________________

IV **Language(s)**

9. Your **first** language **is** ________________________

10. Your dominant language at home is: ________________________

11. Your dominant language at the workplace (or university) is: ________________________

V. Educational History

Please list all qualifications, whether they are ESL-related or not.

	Level of Education	Degree/Diploma/ Certificate	Subject area	Language of education
12.	Secondary school:	____________	____________	____________
13.	Undergraduate:	____________	____________	____________
14.	Postgraduate:	____________	____________	____________
15.	Professional Certification:	____________	____________	____________

VI. Professional Writing Experience

Please characterize, in two or three brief statements, yourprofessional experience in the following areas. lndicate publications, if appropriate, as well as languages other than English used in your professional activities.

16. Writing

____

____

____

____

____

____

17. Editing

____

____

____

____

____

____

18. Other (e.g., Translating)

____

____

____

____

____

VII Experiences Teaching Writing

Please list under the following headings your three most significant teaching experiences:

	Institutional contert	Language(s)	Years
19.	____________________	__________	____
	____________________	__________	____
20.	____________________	__________	____
	____________________	__________	____
21.	____________________	__________	____
	____________________	__________	____

VIII Language Assessment Experiences:

Please list under the following headings your three most significant assessment experiences:

	Institutional context	Skill assessed	Language(s)	Years
22.	______________	__________	__________	____
	______________	__________	__________	____
23.	______________	__________	__________	____
	______________	__________	__________	____
24.	______________	__________	__________	____
	______________	__________	__________	____

25. How would you describe your own skill in assessing ESL writing?
______ Expert ______ Competent ______ Novice

26. How many years' experience do you have in assessing ESL writing?
≤ 2 ____ 3-4 ____ 5-6 ____ ≥ 7 ____

27. Have you taken, or given, a training course in assessing language performance? If so, please describe that briefly.

____

____

____

____

____

____

Appendix B

Interview schedule (semi-structured format)

PART I

What 1 was like you to do is take me through the [12 protocols chosen for prompted recall protocols] one by one and comment on them in light of your background, as a learner, teacher and assessor of English, or anything else that you consider relevant to what you were saying in those protocols.

Part II

Discuss the following aspects of your assessment session in light of your background:

Reading strategy applied to the corpus of compositions

Reading strategy applied to individual compositions

lnterpretation of the role of essay prompts in writing assessments

Performance expectations articulated in the concurrent verbal protocols

Scoring criteria discussed in the concurrent verbal protocols

Use of norm-referencing evident from the concurrent verbal protocols

Attitude displayed towards the writers of the compositions

Would your behaviour have been different if:

You had been told to use a specifíc scoring rubric?

If your assessment had had practical consequences?

Thank you very much for taking the time to answer these questions