TOEFL®

Monograph Series

MS - 22
DECEMBER 2001

Scoring TOEFL Essays and TOEFL 2000 Prototype Writing Tasks: An Investigation Into Raters' Decision Making and Development of a Preliminary Analytic Framework

Alister Cumming
Robert Kantor
Donald E. Powers

ETS® Educational Testing Service

Scoring TOEFL Essays and TOEFL 2000 Prototype Writing Tasks: An Investigation Into Raters' Decision Making and Development of a Preliminary Analytic Framework

Alister Cumming
Robert Kantor
Donald E. Powers

Educational Testing Service
Princeton, New Jersey
RM-01-04

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

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

Email: toefl@ets.org

Web site: http://www.toefl.org

Foreword

The TOEFL Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language™ (TOEFL®) test development efforts. As part of the foundation for the TOEFL 2000 project, a number of papers and reports were commissioned from experts within the fields of measurement and language teaching and testing. The resulting critical reviews and expert opinions have helped to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project is a broad effort under which language testing at Educational Testing Service® (ETS®) will evolve into the 21st century. As a first step, the TOEFL program recently revised the Test of Spoken English™ (TSE®) and introduced a computer-based version of the TOEFL test. The revised TSE test, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The computer-based TOEFL test, introduced in 1998, takes advantage of new forms of assessment and improved services made possible by computer-based testing, while also moving the program toward its longer-range goals, which include:

- the development of a conceptual framework that takes into account models of communicative competence
- a research agenda that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

Monographs 16 through 20 were the working papers that laid out the TOEFL 2000 conceptual frameworks with their accompanying research agendas. The initial framework document, Monograph 16, described the process by which the project was to move from identifying the test domain to building an empirically based interpretation of test scores. The subsequent framework documents, Monographs 17-20, extended the conceptual frameworks to the domains of reading, writing, listening, and speaking (both as independent and interdependent domains). These conceptual frameworks guided the research and prototyping studies described in subsequent monographs that resulted in the final test model.

As TOEFL 2000 projects are completed, monographs and research reports will continue to be released and public review of project work invited.

TOEFL Program Office
Educational Testing Service

Abstract

This project established a framework to describe the decision-making processes that experienced writing assessors use to evaluate compositions written by students who speak English as a second language (ESL) or as a foreign language (EFL). The framework will assist in the development and field testing of a scoring scheme for the writing component of the new Test of English as a Foreign Language (TOEFL 2000) examination. Phase One developed empirically a) an initial framework to describe the decision-making behaviors of 10 experienced ESL/EFL instructors/assessors of differing backgrounds, each of whom produced concurrent verbal reports of their decision making while rating 60 TOEFL essays and b) a questionnaire to profile relevant variables in the raters' backgrounds. Phase Two refined the framework, gathering additional think-aloud data from a) seven highly experienced native-English-composition assessors while they rated 40 of the Phase One TOEFL essays and b) seven of the Phase One ESL/EFL instructors/assessors while they rated six TOEFL essays and 30 ESL compositions written for five TOEFL 2000 prototype tasks that involve responding to reading or listening materials. Phase Two analyses confirmed the utility of the descriptive framework and produced recommendations for the development of scoring procedures for writing tasks that integrate reading or listening materials.

Key words: Assessment of ESL writing, rating processes, think-aloud verbal reports, TOEFL, background influences on assessment, evaluating compositions written from sources

Acknowledgments

We gratefully acknowledge assistance from the 10 people who, as graduate-student research assistants at the Ontario Institute for Studies in Education of the University of Toronto, played an integral role in all stages of this research: Lindsay Brooks, Michael Busch, Usman Erdosy, Hameed Esmaeili, Irena Ganeva, Mark James, Toshiyo Nabei, Nancy Reiner, Gail Stewart, and Anthony Tong. We are grateful also to the seven U.S. raters who assessed compositions in Phase Two of the study. Carol Taylor and Terry Santos deserve many thanks, as well, for their advice and encouragement while working with us on this and other research to prepare a new writing component for the Test of English as a Foreign Language (TOEFL 2000) examination. We also thank several anonymous reviewers of earlier versions of this report for helping us to clarify various points in the manuscript, as well as Marie Collins for her careful copy editing of the final version of this report.

Table of Contents

List of Tables

1. Overview

The purpose of this project was to develop, then refine, an analytic framework that describes how experienced writing assessors evaluate compositions written by individuals who speak English as a second language (ESL) and as a foreign language (EFL). This descriptive framework is intended to serve as a basis for the development and field testing of a scoring scheme for the writing component of a new version of the Test of English as a Foreign Language (TOEFL 2000). This report documents Phase One and Phase Two of the two-part study.

Phase One, which was conducted from June 1998 to August 1998, focused on developing empirically a) an initial framework to describe the decision-making behaviors of 10 experienced ESL/EFL instructors/assessors and b) a questionnaire to profile individual characteristics of the raters as well as relevant variables in their backgrounds. To develop the descriptive framework, the participating assessors — who had diverse backgrounds — produced verbal reports on audio tape while rating 60 TOEFL essays in Toronto. Phase Two, which was conducted from September 1998 to December 1998, refined the descriptive framework and the questionnaire developed in Phase One. In Phase Two, we gathered additional think-aloud data from seven experienced native-English-composition assessors in various parts of the United States as they rated 40 TOEFL essays, as well as from seven of the same experienced Phase One ESL/EFL instructors/assessors in Toronto (with one change in personnel) as they rated five TOEFL 2000 prototype tasks piloted by six ESL students in California. In Phase Two, we also coded and analyzed matched samples of the think-aloud data from both phases of the project, and we interviewed participants about their impressions of the prototype tasks, asking for their suggestions for improvement.

The following findings from Phase One are presented in this report:

- influences raters perceived on their assessment performance (a draft survey instrument used to profile raters' characteristics is provided)
- a preliminary framework to describe the decisions raters most frequently made while assessing TOEFL essays
- typical decision-making sequences evident in the data
- text features in the TOEFL essays to which the raters attended
- impressions of variations in these behaviors by levels of English proficiency represented in the compositions, by essay topic, and by characteristics of individual raters

The following findings from Phase Two are presented in this report:

- influences raters perceived on their assessment performance
- trends in decision making by native-English-composition raters
- trends in decision making for TOEFL 2000 prototype tasks
- refinements made to the descriptive framework of decision-making behaviors
- statistical descriptions of coded think-aloud data
- formative assessment of TOEFL 2000 prototype tasks

In addition to documenting raters' thinking processes in detail, we found both groups of raters used fundamentally similar decision-making behaviors, in similar proportions of frequency, while assessing both TOEFL essays and prototype writing tasks — thus verifying our descriptive framework. On the basis of these findings, we make the following recommendations:

1. to use the descriptive framework to design a project in the near future that uses the present findings to prepare and field test scoring rubrics and procedures
2. to monitor background influences on assessors' scoring behaviors
3. to continue to develop prototype writing tasks that integrate reading and listening, specifying carefully the instructions in the tasks, criteria for judging them, and the uses that examinees are expected to make of particular source materials and genres in their writing

2. Relevance to TOEFL 2000 Project

As argued in Cumming, Kantor, Powers, Santos, and Taylor (2000), revisions planned for the TOEFL 2000 examination require a new, valid method for scoring a variety of types of written compositions. Although the holistic scoring scheme now utilized for TOEFL essays (formerly the Test of Written English, or TWE®) can serve as a well founded benchmark to begin this process, it lacks the descriptive model and research base needed to ensure confidence in what its scores mean in terms of student abilities. Moreover, the TWE rating scale cannot simply be extended, without extensive research, to the other task types envisioned for the writing component of the TOEFL 2000 test. Also, the information it provides for score users is too limited for the purposes of the TOEFL 2000 test. In short, a new scoring scheme for TOEFL 2000 writing tasks needs to be based on close analyses of raters' decision-making processes, to be applicable to several types of writing, to be validated, and to be descriptively informative for reporting purposes.

Cumming et al. (2000) proposed that a "reader-writer model" is necessary for the writing component of the TOEFL 2000 exam. The present project takes the first steps in this direction. Before subsequent steps in the development of TOEFL 2000 writing tasks can be taken, a) new prototype tasks must be field tested, b) alternative scoring schemes must be considered and their utilization with raters evaluated, and c) the difficulty and other properties of writing tasks must be determined. These latter steps will be undertaken in projects subsequent to the present one, but the current study is a necessary, preliminary step. That is, a descriptive framework must be developed and validated that describes raters' decision-making behaviors when scoring a range of writing tasks relevant to the aims of the TOEFL 2000 project *before* research can be done to select an appropriate scoring method, to field test prototype writing tasks, to determine the relative difficulty of those tasks, and to know how to train raters of compositions effectively on such writing tasks.

Need for the research conducted in the present study is evident in numerous recent publications that have questioned the construct validity of existing holistic or other types of analytic schemes for scoring written compositions, both in first-language contexts (e.g., Charney, 1984; Huot, 1990; Purves, 1992) and second-language contexts (e.g., Connor-Linton, 1995; Cumming, 1997; Hamp-Lyons & Kroll, 1997; Kroll, 1998; Raimes, 1990). These analyses echo doubts that many others in relevant educational communities have expressed as well. Many holistic schemes for scoring writing can, with extensive rater training and monitoring for example (e.g., Stansfield & Ross, 1988; Weigle, 1994), produce reliable, consistent assessments; but a principal criticism has been that the exact nature of the constructs they assess remains uncertain. For instance, holistic rating scales can conflate many of the complex traits and variables that human judges of students' written compositions perceive (such as fine points of discourse coherence, grammar, lexical usage, or presentation of ideas) into a few simple scale points, rendering the meaning or significance of the judges' assessments in a form that many feel is either superficial or difficult to interpret (Henning, 1991; Purves, 1992; Raimes, 1990). The simplicity of the holistic scoring method, and the rating scales that typically accompany it,

obscures its principal virtue: reliance on the complex, richly informed judgments of skilled human raters to interpret the quality of students' writing abilities.

Importantly, there has been little research to describe with precision what skilled human raters attend to when they score compositions, particularly how such assessments correspond to students' actual proficiency in writing in a second language. As a consequence, many holistic and other analytic rating scales lack firm empirical substantiation with respect to evidence about second-language learners' writing abilities. Indeed, the methods typically used to develop holistic scoring rubrics — for example, sorting students' essays into levels based on judgments of quality, and then identifying characteristics in students' texts that distinguish among the levels (e.g., Ruth & Murphy, 1988) — have produced rating scales that are too imprecise and broadly defined to demarcate individual students' achievements in writing, either for the purposes of achievement testing or of research into second-language writing development (Cumming, 1997; Polio, 1997). Moreover, there is considerable debate over which types of holistic scoring methods are most appropriate and informative for second-language assessment (Cumming, 1997; Hamp-Lyons, 1991):

- truly holistic scoring schemes that combine several traits into single score points?
- multitrait scoring methods that assign ratings to distinct aspects of compositions (e.g., discourse organization, presentation of ideas, grammar, and vocabulary)?
- primary trait methods that assign ratings to key features of compositions expected for a specific writing task (Lloyd-Jones, 1977)?

Two further issues are a) the variability in cultural standards that implicitly inform holistic scoring and b) the sampling of types of written genres for assessment purposes. Numerous recent studies have found groups of raters to differ in their evaluations of writing according to their cultural or disciplinary backgrounds and according to the types of written genres assessed (e.g., Kobayashi & Rinnert, 1996; Mendelsohn & Cumming, 1987; Song & Caruso, 1996). More information is needed about the range of scoring behaviors that experienced raters with diverse academic, cultural, and experiential backgrounds demonstrate when using holistic scoring. Such information is needed to identify differences as well as commonalities among experienced raters and also to determine which rating behaviors might be worth either promoting or discouraging in the context of particular writing assessments. Similarly, information is needed to describe how raters' scoring behaviors may vary across different types of writing tasks. Many of the schemes now widely used for holistic scoring of adult ESL/EFL learners' compositions (e.g., the current rubric for TOEFL essays; Hamp-Lyons & Henning, 1991; Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey, 1981) were devised only to rate brief, argumentative-type essays; they are not suitable for scoring other types of writing tasks, for example, that might involve the integration of content from reading or listening sources.

For these reasons, we started a program of intensive research into the decision-making behaviors that experienced raters of compositions actually use when they evaluate ESL/EFL students' compositions. Such research has been proposed for over a decade in respect to first-language writing (e.g., Freedman & Calfee, 1983; Huot, 1990, 1993; and see DeRemer, 1998, for a recent study), but has been taken up in only a few, isolated, exploratory studies on second-language writing (e.g., Connor & Carrell, 1993; Cumming, 1990; Milanovic, Saville, & Shuhong, 1996; Sakyi, 2000; Vaughn, 1991). At present, there is to our knowledge no well substantiated model that accurately describes the rating behaviors of skilled evaluators of ESL/EFL compositions. Such a model is needed a) to begin to formulate, field test, and validate appropriate scoring schemes for ESL/EFL composition assessment; b) to guide and monitor the training of raters in the use of these scoring schemes; and c) to report results to score users in a valid, informed manner.

We conducted this study in two parts. In Phase One we developed the preliminary basis for such a model, then in Phase Two we refined and evaluated it. Although the current study was exploratory, its grounded basis in analyses of raters' decision making, and its progressive sequence for expanding and refining the descriptive model directly in relation to the proposed writing tasks for the TOEFL 2000 examination, have produced results capable of informing further decisions about the scoring scheme for the revised assessment.

In this paper, we report the goals, methods, and findings of the two phases of this project separately, followed by recommendations for further research and development. Our research followed an emergent sequence: its analyses are grounded on a progressive sequence of empirical studies, conducted in a collaborative mode, in which the participating graduate students and university professor generated, as well as collectively analyzed, their own and each other's behaviors while rating ESL/EFL compositions. Rather than imposing or creating ad hoc, a framework for scoring TOEFL writing, the project aimed to construct and progressively refine an analytic framework from close examination of raters' behaviors while scoring samples of ESL/EFL compositions relevant to TOEFL 2000 goals and objectives. The value of this approach is that the research team was able to share and verify interpretations of the data and to develop a collective understanding of them as a whole — in a way that would not have been possible if the roles of assessors and researchers had been isolated and separated from each other.

To counter the possible problems of insularity, bias, and idiosyncrasy this approach may have produced, Phase Two of the project employed external raters in the United States, each of whom is highly regarded for his or her expertise in evaluating native-English compositions for Educational Testing Service (ETS), such as those written for the SAT® II: Writing Test and the Graduate Record Examinations® (GRE®) Writing Assessment. We analyzed the performance of these raters in relation to that of the research team in Toronto, which assessed comparable sets of TOEFL essays as well as TOEFL 2000 prototype tasks that involved reading and listening materials. Collectively, the raters participating in this research were diverse — linguistically and experientially — including people with extensive experience in settings where English is a

second language, a foreign language, and a native language. In examining the performance of a broad range of raters, we hoped to account for some of the diversity in standards and experience under which English is assessed internationally, as well as to identify commonalities within that diversity.

Both phases of the study made extensive use of think-aloud verbal reports made by raters as they scored ESL/EFL compositions. This method of inquiry has proved illuminating in previous studies of the decision-making processes involved in ESL/EFL composition evaluation (e.g., Cumming, 1990; Vaughn, 1991). Indeed, many have preferred this approach generally for research into composition assessment (e.g., Huot, 1990; Wolfe, Kao, & Ranney, 1998) and for other inquiries into human performance on complex tasks (e.g., Ericsson & Simon, 1984), including many aspects of second-language research (e.g., Cohen, 1994). Care was taken, both when instructing and training raters to generate the think-aloud protocols and when interpreting research results, a) to acknowledge that such data represent only information that people attend to during task performance (i.e., not their full cognitive processes, which probably cannot be verbalized), and b) to avoid or acknowledge problems of reactivity or other inadvertent influences on participants' behaviors (cf. problems identified in various chapters in Smagorinsky, 1994).

3. Phase One: An Initial Framework

Introduction

Phase One of this project addressed the principal question below, along with four related subquestions:

What aspects of TOEFL essays do experienced instructors/assessors of ESL/EFL compositions attend to while they assess a purposive sample of these essays (demarcated by English proficiency level and topic differences)?

1. What are the principal characteristics of the instructors/assessors participating in the study, such as might bear on their rating behaviors?

2. What are the integral decisions the instructors/assessors make to monitor their rating behaviors while rating, as documented in concurrent verbal reports of their thinking while assessing the essays?

3. Which text features of the essays do the instructors/assessors attend to for assessment purposes, as documented in concurrent verbal reports of their thinking while assessing them?

4. Do these raters' decisions and the text features they perceive vary appreciably with the examinee's level of English proficiency (as determined by previous TOEFL ratings), with the essay topic, with particular reader characteristics, or with individual raters?

Method

Ten experienced ESL/EFL instructors were selected for the diversity and extent of their experiences internationally — nine graduate students and one professor at the University of Toronto. Each rated randomly sequenced samples of 60 TOEFL essays, which were selected by ETS staff to represent a range of four essay topics and six score points from a recent version of the test. All raters produced concurrent think-aloud protocols to document their decision-making behaviors while they rated the compositions. Audio tapes of these verbal reports were transcribed by the research team. The participants who had generated the verbal reports then verified and corrected the transcripts. Next, we analyzed the data to address the four subquestions guiding this phase of the project. Each rater also completed a profile questionnaire — which had earlier been developed, pilot tested, and refined by the research team — to document relevant individual characteristics and experiences, such as may relate to their rating behaviors. Analyses were primarily impressionistic at this initial stage of the study.

Participants

The 10 participating raters — nine graduate students and one professor — were selected based on their extensive experience teaching and assessing ESL/EFL writing, the diversity of their backgrounds and experiences internationally, prior completion of graduate courses in language assessment and research methods, and their availability for part-time work in Toronto during the time of the research. All participants acted both as "subjects" in the research and as paid "researchers," designing and refining instruments as well as generating, transcribing, verifying, and analyzing data. Co-investigators at ETS collaborated on the design of the proposal, selected sample compositions for rating, provided advice during the research, and offered input into a draft of this report.

To preserve the confidentiality of individual raters, data on individuals are reported in aggregate here. All participants opted for pseudonyms that preserve their gender but conceal their ethnicity and other identifying characteristics: Ed, Gary, Jane, Karen, Marty, Melissa, Patricia, Paul, Roy, Scott, and Zoey (a total of 11, because the pseudonym of one additional rater, who participated in Phase Two only, is included to keep that individual's identity confidential). All participants who are not ETS employees signed forms approved by ETS' Prior Review Committee, which indicated their informed consent in view of ethical considerations in the research; they also signed agreements to preserve the confidentiality of all information related to the project, not to make any uses of it in their work, and to surrender any claims to the products of the research. The research team met weekly (for most weeks) between June and August 1998, reporting on and discussing ongoing aspects of the project's work, which was conducted mostly in pairs or individually.

According to information the raters provided by way of the profile questionnaire (see Appendix A) the 10 participants had the following characteristics:

- They were in their late 20s (one person), 30s (four people), or 40s (five people).
- The group was balanced in their gender distribution: five males and five females.
- They had a range of types of experience teaching/assessing ESL/EFL to adults, and in a few cases to high school students, in academic contexts (three in ESL contexts in North America, and seven in both ESL contexts and, more extensively, EFL contexts in Bulgaria, Hong Kong, Iran, Portugal, or Japan).
- Each had taught/assessed ESL/EFL for periods of from two to 19 years (six for more than seven years, three for three to five years, and one for two years).

- All had completed a Teaching English as a Second Language (TESL) or related teaching certification program.

- All a) held a Ph.D. (two people), b) were midway through a Ph.D. program in second language education (four people), or c) were just completing (two people) or midway through (three people, one of whom held a Ph.D. in another discipline) a master's degree in second language education.

- They were native speakers of either North American English (five people) or of Bulgarian, Cantonese, Hungarian, Farsi, or Japanese. Each of the latter had either resided in Canada for much of their adult lives (two people), or had passed the university's TOEFL-score entry requirement of at least 580 and previously completed at least one year of graduate study at the university successfully, all with average grades of A. In addition, all routinely used English in their work and graduate studies, and were sufficiently proficient in English to work part-time in Canada as ESL instructors.

- They rated their ESL/EFL composition-rating abilities as either "expert" (three people who had extensive experience as raters for university ESL composition exams and who trained others on composition assessment), "competent" (six people who had several years' experience rating university or high school composition exams, two of whom had also trained teachers in assessment), or "novice" (one person with little such experience, other than tutoring university ESL students for two years).

- They had a range of previous experience writing professionally (four people with extensive publications, the other six with some professional publications, mainly in educational contexts), editing work for publication (five people), and/or translating (three people).

Approach

Profile questionnaire. The research team first developed a questionnaire to profile the characteristics and experiences of raters, such as might bear upon their rating behaviors. We then revised it four times by piloting it on ourselves. A draft version of this questionnaire was completed individually by all study participants immediately or shortly after generating their think-aloud protocols. The version of the questionnaire included in Appendix A was refined slightly after its administration, based on feedback from the research team in this phase of the research.

Sample essays. ETS staff selected a pool of 143 TOEFL essays from a recent administration of the test, then conveyed these to Toronto, blinded as to the names, locations, or other identifying characteristics of examinees. The essays represented four different topics[1] and the full range of six possible score points. About half were written by hand, and about half were typed. We assigned new code numbers to the compositions and, when photocopying them, concealed the original scores they had been assigned. We used Minitab, Inc.'s (1997) program for assigning random numbers to generate a unique random sequence of 60 of the compositions for each rater, sampling randomly from the whole pool of 143 compositions. Randomization was done to counteract the effects of sequencing the compositions on raters' behaviors, and so that the participants would not try to predict how many compositions at particular score points would be assigned to them.

Think-aloud protocols. All raters received an initial demonstration of the think-aloud process. Instead of essays, training in thinking aloud involved mathematics problems and counting windows in houses (following Ericsson & Simon, 1984), so as not to bias raters' behaviors for rating the compositions. After the demonstration, raters practiced the procedure in pairs for about half an hour with several additional math problems, and then discussed the procedure collectively. As orientation, each participant read a copy of the proposal for the present project, and read Cumming (1990) as an example of previous, related research. The research team generated and collectively revised a standard set of instructions for generating the think-aloud protocols (see Appendix B).

Rating process. In early July, each participant received a package of 60 randomly sequenced TOEFL essays (in a unique order for each rater), a copy of the four essay prompts in response to which the compositions had been written, a tape recorder, several cassette tapes, and instructions. The raters were asked to assign a number from 1 to 6 for each composition — with 1 representing least proficient writing in English and 6 representing most proficient writing in English — while simultaneously generating and recording a verbal report of all of the things they were thinking about while performing this task. To contextualize the assessment task and to elicit information about each rater's own conceptualizations of differences in ESL/EFL writing quality, each participant was asked to assign a rating to each composition in the manner of a q-sort methodology. That is, they were asked to judge the quality of the compositions without our specifying in advance which criteria to use. They were also asked to make their evaluations without direct reference to the scale for scoring TOEFL essays (which none of the raters had used in their evaluation practices, though a few were familiar with it). In other words, the participants were left to their own devices and judgments as to how they should rate the compositions. They completed this task in the convenience of their homes, typically within 1 or 2 days, then completed the profile questionnaire.

[1] For reasons of test security, the specific essay prompts are not reported here.

Transcription process. The research team created and revised a standard set of simple transcription conventions (see Appendix C). Each tape-recorded think-aloud protocol was transcribed by a member of the research team who had not produced the particular think-aloud protocol. Afterward, to assure accuracy, the transcriptions were verified and corrected by the person who had originally produced the verbal report. Once the transcriptions were complete, copies were distributed to pairs of researchers, who then conducted preliminary analyses to address the subquestions guiding this phase of the study. The think-aloud protocols were extensive, ranging in length from 21 to 56 typed pages per rater, or between 1/6 of a page and two pages per composition.

Analyses. For preliminary analyses, pairs of researchers reviewed all of the transcribed data; generated tentative, impressionistic interpretations; then confirmed and discussed these with each other, checking further on samples of the transcribed data for confirmation. Afterward, they reported their interpretations to the full research team, which further discussed and refined the interpretations. Hence, the findings reported in the next section are primarily impressionistic and qualitative, though grounded in a thorough review of the entire data set, as well as confirmed by subsequent checks on the data — not only by the participants who produced the data but also by the entire research team.

Much of the discussion among the research team concerned the terminology we should use to describe raters' behaviors and the criteria they displayed, with the aim of reaching a consensus on phrases that could most faithfully convey these complex concepts. Originally, we had intended that Phase One would involve coding of data; however, all members of the research team quickly realized in viewing the transcribed data that simply preparing numerical tallies of particular aspects of the verbal reports would not produce meaningful analyses at this early stage of the inquiry. As noted earlier, the think-aloud protocols produced verbal reports of the aspects of the compositions that the raters reported attending to, rather than full reports of their decision-making or other cognitive processes. Although the verbal reports were relatively full and extensive, it was evident that for any particular composition, a rater's verbal reports represented only a fragment of the aspects of the compositions that the person attended to while reading and assessing the compositions. The reports were often just glimpses of the criteria people actually used to guide their scoring. For this reason, our analyses have treated the verbal reports and the research participants as a whole, primarily seeking impressionistic interpretations of patterns and trends in the data overall. This limitation is countered by the sheer extent of the data collected (which provides ample information on each individual rater's approach to assessing the compositions), the overall diversity in the participants' backgrounds (which provides a wide range of perspectives and orientations to this kind of assessment), and the focus in Phase One of the study on generating initial descriptions of raters' decision-making behaviors.

A primary purpose of this research was to find out what criteria and procedures the raters would utilize. Thus, each rater was encouraged to use criteria and procedures of his or her own devising (rather than the scale for scoring TOEFL essays or any other formal rating scheme). Since the raters' scores do not derive from uniform rating criteria (and thus could not logically be comparable), we have not analyzed the actual ratings that were assigned to the compositions. Moreover, because each rater received a unique, randomly assigned set of compositions to evaluate (that is, the ratings were not done on a common set of compositions), the number of compositions rated in common among the raters was too few, across the six score points, to permit us to conduct inferential statistics on the data arising from this study. As described in the next section, only 20 of the 143 TOEFL essays were rated by two or more of the 10 raters. Among these, only seven essays were rated by at least six of the raters, and the distribution of most of these seven was skewed by the randomization program at the two lowest points on the six-point rating scale.

Findings

Findings from Phase One of this study are reported in terms of the four subquestions that guided this phase of the project.

Characteristics of Participants Influencing Their Ratings

Research subquestion 1 asked what the principal characteristics of participants in the study were, such as might bear on their rating behaviors. The principal characteristics of the raters participating in this phase of the study were summarized earlier, under the heading Participants. These profile data make evident the diversity in participants' backgrounds as well as their common, extensive experience in teaching and assessing adult ESL/EFL writing.

Past influences on rating behaviors. Three items in the profile questionnaire (see Appendix A, section I, items 1 to 3) asked participants to document and assess the influences, from past experiences, that they thought affected their rating behaviors in the present task. On a scale from 1 to 5, where 1 = "not at all," 3 = "slightly," and 5 = "a great deal," the raters reported the following:

- Two raters thought there were no such influences on their ratings. One of these two raters stated he had not taught composition for two years, and therefore had "forgotten the scales [he] had used before," while the other rater did not elaborate on this point.

- Three raters stated there were very slight influences (self-ratings of 2) on their ratings. One of these raters stated she was "familiar with both TWE and [another Canadian rating scale] bands, and they might have influenced me to a slight degree." One

stated she knew only about TOEFL essays or TWE "in a general sense, but not a specific way." And the other rater indicated she was not influenced on any "specifics apart from the anchors of 1 as incompetent and 6 as native-like competence."

- Three raters thought their ratings were "slightly" influenced (self-ratings of 3) by previous rating scales they had used. Each person cited particular but different rating scales they had used extensively in their previous work or research: one from a university in Canada, one from a published, widely used scale for assessing writing, and one from a high school assessment system in another country.

- One rater thought he was influenced (self-rating of 4) by a rating scale he had used extensively in his work as a rater and instructor at a Canadian university.

- One rater stated he was influenced "a great deal" (self-rating of 5) by several published and widely used schemes for assessing ESL writing as well as a marking scheme used at the university where he had worked (in a country other than Canada).

In sum, most of the participants believed that their ratings had been influenced to at least some extent by their previous experiences as evaluators or instructors of ESL/EFL. The perceived extent of this influence varied greatly, however, as did the range of possible influences cited. These influences included a variety of published scales commonly used to assess ESL/EFL composition (e.g., Hamp-Lyons & Henning, 1991; Jacobs et al., 1981; the scale for scoring TOEFL essays), scales used at particular universities or in high school systems, and unique rating scales certain people had used in their teaching.

This finding is not surprising, given a) that the effects of training on raters of ESL/EFL composition are well documented (e.g., Wiegle, 1994), b) that all participants have devoted their careers to second-language education and assessment, and c) that the nature of the overall task was very open-ended (i.e., raters were encouraged to devise their own scoring schemes). But the implication of the finding for the TOEFL 2000 project, and for further TOEFL 2000 research, is that raters of ESL/EFL composition will probably be influenced, to varying extents, by scales and criteria they may have used previously to assess ESL/EFL compositions. Indeed, it may be relatively difficult for experienced raters to "unlearn" or shift themselves away from particular criteria or procedures that they have become skilled at using. Erdosy (2000), in subanalyses of these and other data from Phase One, provides further evidence that the raters were influenced primarily by their previous experiences teaching and assessing ESL/EFL.

There was a pattern in these self-report data (again not unexpected) that the more experience raters had with specific scales for rating ESL/EFL compositions, the more they perceived themselves to be influenced by such scales. This trend was not consistent, though. Two of the most experienced raters (both of whom described themselves as "expert" raters of ESL

compositions) rated their influences as only "slight" (ratings of 2 or 3), perhaps as a result of their having previously used such a wide range of schemes for composition assessment that they were not tied to any particular scoring scheme. Nonetheless, it can be presumed that the development of wholly unique scales for assessing ESL/EFL writing for the TOEFL 2000 examination is inevitably going to be shaped (or constrained) by the prior history and influences of existing schemes and practices, such as now are widely in use. But if the TOEFL 2000 test is going to reflect prevailing standards for ESL/EFL writing ability, how could things be otherwise?

Elements of effective writing. Another item on the profile questionnaire (see Appendix A, section I, item 4) asked participants "what three qualities [they] believe make for especially effective writing in the context of a composition examination?" Participants used various terms to respond to this question, the most frequently mentioned of which related to:

- rhetorical organization, including introductory statements, development, cohesion, and fulfillment of the writing task (mentioned by nine raters)
- expression of ideas, including logic, argumentation, clarity, uniqueness, and supporting points (mentioned by nine raters)
- accuracy and fluency of English grammar and vocabulary (mentioned by seven raters, with two raters citing grammar and vocabulary as separate categories)
- the amount of written text produced (mentioned by two raters)

The brief statements about these text qualities that participants wrote on the questionnaire could hardly do justice to the complexity of concepts that these terms may represent. Nor do they illuminate the differing perspectives that individual raters may have had on these terms in view of specific text qualities or text types (as evidenced by research on contrastive rhetoric, for example). These terms may be conventional and common to ESL/EFL assessment practices; there was some uniformity in participants' mentioning of them, suggesting they may point toward important categories (or traits) in scoring English writing. But people may associate different values and interpretations with these terms in specific contexts or in regard to different genres of writing. Therefore, we think it important to look to the criteria that emerged from the think-aloud data to determine more precisely (than is possible in written questionnaire responses) the criteria that these raters used to make judgments about writing quality in this context.

The only trends between rating behaviors and questionnaire responses that we noticed at this stage related to participants' previous teaching, assessing, writing, editing, or translating experiences. More experienced raters seemed to verbalize more extensively their assessment decisions (in the think-aloud protocols, some of them produced longer and more verbose protocols) and their experiences (similarly, some produced more detailed questionnaire

responses). Also, in some cases more experienced raters tended to adopt certain specific macrostrategies for comparing their ratings of the compositions (such as reviewing them collectively or sorting them into piles upon initial inspection; these are discussed in more depth in a later section). These impressions echo similar findings in Huot's (1993, p. 226) comparative study of expert and novice assessors of native-English compositions. However, the selection of participants for this phase of the study was not designed to evaluate such influences systematically, so we did not undertake detailed analyses of these variables.

Decision-Making Behaviors While Rating

To answer subquestion 2 — *What are the integral decisions the instructors/assessors make to monitor their rating behaviors while rating, as documented in concurrent verbal reports of their thinking while assessing the essays?* — we analyzed the decision-making behaviors displayed in the think-aloud protocols two ways. First, we considered the schemes developed by Cumming (1990) and Sakyi (2000), presented in Appendix D, as preliminary frameworks to describe decision-making behaviors while rating ESL/EFL compositions; we then assessed the adequacy of these schemes to account for the present data. In particular, we searched for decision-making behaviors that might not have been identified in the earlier research (i.e., in either Cumming, 1990, or Sakyi, 2000) or that may not have been described in a way that accurately reflected the present data. This first analysis identified specific *types* of decisions that the raters, as a whole, made while rating the compositions. The second analysis considered the *sequences* of decision making that they made.

Our initial inspections of the think-aloud data verified that all of the categories of behavior described in Appendix D appeared frequently in the present data. However, certain other decision-making behaviors also appeared in the present data — seemingly because the present data contained a wider range of levels of English proficiency and because we used raters with differing backgrounds (more so than in the studies by Cumming, 1990, or Sakyi, 2000). As a result, we sought to verify, extend, and refine this framework by matching the descriptive terms in the preliminary framework to the present sets of think-aloud protocols. The major points of verification of the earlier framework were:

- We reaffirmed the basic distinction between *interpretation strategies* (reading strategies aimed at comprehending the composition) and *judgment strategies* (evaluation strategies for formulating a score). This distinction seemed evident in data from all 10 raters, forming "a sort of tension between the reader as reader and reader as rater," as Huot, in analyzing the rating behavior of native-English-composition assessors, termed it (1990, p. 255).

- We reaffirmed the three general categories of decision-making behaviors, with a focus either on *self-monitoring* of one's own rating behaviors, the composition's realization of *rhetorical and ideational* elements, or the composition's accuracy and fluency in the English *language*.

- We verified the presence of all of the decision-making behaviors initially described in either Cumming (1990) or Sakyi (2000) in each of the sets of think-aloud data from each of the 10 raters in the present study.

The major alterations we made to the earlier descriptive framework to account for the present data are described as follows:

- We added the behavior *decide on a macrostrategy for reading and rating* the compositions to account for the ways in which several of our raters approached the overall assessment task differently. Some either just read the compositions and rated them, while others sorted the compositions into piles, according to their first impressions of the compositions, prior to assessing them more thoroughly.

- We added the behavior *assess writing skill* to account for several raters' frequent comments about the literacy, writing expertise, lack of experience, and so forth of the examinees.

- We added the behavior *assess style, register, or genre* to account for raters' frequent remarks about the style of writing — informal or formal, colloquial or conversational, academic or nonacademic, and so on.

- We added the behavior *assess reasoning or logic* to account for the frequent attention the raters paid to the quality of ideas expressed, their logic (or lack of logic), and qualities such as presumptuousness, triteness, unusualness, and so on.

- We added the behavior *assess task completion* to account for frequent remarks in the data about whether the examinee had, or had not, successfully completed the task assigned in the essay prompt.

- We added the behavior *assess fluency* to account for this criterion being frequently mentioned by all raters.

- We added the behavior *monitor for personal biases* to describe how some raters checked whether their own preferences, past experiences, or viewpoints might be inadvertently affecting their assessments.

- We revised the wording describing most of the behaviors to more precisely describe what we thought appeared in the present data (e.g., changing "establish" to "consider," adding phrases, etc.).

Table 1 displays the preliminary descriptive framework after all verifications and revisions — a total of 37 distinct decision-making behaviors that appeared frequently in the Phase One think-aloud data. Comparing the behaviors listed in Table 1 to related lists or schemes produced by other research into the decision-making behaviors of composition assessors (e.g., DeRemer, 1998, pp. 23-24; Pula & Huot, 1993, pp. 243-244; Vaughn, 1991, pp. 119, 122-125; Wolfe, Kao, & Ranney, 1998, pp. 472-477), we are confident that this framework, in addition to accounting for our own data comprehensively, also encompasses the most integral decision-making behaviors identified in these other studies (though from a different analytic and contextual perspective). Similarly, the behaviors identified here correspond, though only generally, to those identified by the raters as important criteria they said they applied during their assessments (see *Elements of effective writing,* earlier). Nonetheless, this version of the descriptive framework was tentative and preliminary, pending further analyses and refinements reported later for Phase Two of the present study.

Table 1

Preliminary Descriptive Framework of Rater's Decision-Making Behaviors While Scoring TOEFL Essays

Self-monitoring focus	Rhetorical and ideational focus	Language focus
	Interpretation strategies	
* read or reread essay prompt * read or reread composition * envision personal situation and viewpoint of writer	* interpret ambiguous or unclear phrases * discern rhetorical structure * summarize ideas or propositions	* scan whole composition * classify errors into types * edit or correct phrases for interpretation
	Judgment strategies	
* decide on macrostrategy for reading and rating * consider own personal response * define and/or revise own criteria * compare with other compositions or "anchors" * distinguish interactions between categories * summarize and tally judgments collectively * articulate or revise scoring decision * monitor for personal biases	* assess reasoning or logic * assess task completion * assess relevance * assess coherence * assess topic development * assess interest, originality, or creativity * identify redundancies * assess helpfulness in guiding readers * assess style, register, or genre * rate ideas and rhetoric	* assess quantity of total written production * assess comprehensibility * consider gravity of errors * consider error frequency * assess fluency * assess writing skill * consider command of lexis * consider command of syntax and morphology * consider command of spelling and punctuation * rate language overall

Note. Neither Table 1, nor the following versions of Table 1 that are presented later, are intended to indicate any particular sequences in which these decision-making behaviors might occur.

Sequences of Decision Making While Rating: Macrostrategies

In a second analysis of the think-aloud data, which also aimed to address subquestion 2, we identified typical sequences of decision making. Globally, we observed a fundamental

distinction between macrostrategies and microstrategies. As noted earlier, we adopted the term macrostrategies to account for behaviors raters exhibited when deciding how to deal with the overall set of compositions. We had not provided any specific directives about this behavior, because we wanted to see how participants approached this task independently. Most raters simply started reading and assessing the compositions in the sequence in which they had been presented to them. For example:

> And right now I'm looking at the first composition. I think I will just take the compositions in the order in which they come, uh, because, uh, I don't really feel like reordering them. (Scott)

But others spent time deciding how they would first approach the task as a whole. For instance:

> All right, so I think what I'll do is, just to get warmed up is, pull some of the shorter ones, put them at the front and start with the easier ones. So I am just looking through my pile, pulling out some of the ones that I can see that are obviously extremely short and good candidates for getting a score of 1 ... based on their length ... O.K., I am still looking through my pile, trying to identify some of the shorter ones ... I want to start out reading some of the easier ones until I'm familiar, until I feel comfortable rating. O.K., so I've pulled out, uh, I don't know, at least half a dozen of the shorter ones, and I'd like to start with those. (Jane)

> So now I'm sort of making little piles. I'm putting this 62 on my left and I'm putting this, uh, which one was the one that was so good? I'm putting the 88 on my right, and I'm just trying to lay the others in between in the sort of, so I'm putting 7 right in the middle. Well, no. Now I'm going to put it just, I'm going to put it in pile number two. (Zoey)

And others exercised macrostrategies later during their assessments, reviewing all and revising some of their preliminary scores for the compositions. For example:

> So, I've now read them all once, initially, over and I feel I should have a look again, to sort of standardize my scores and reconsider them. Now, I, I don't personally feel, I kind of feel on these tasks, that I, my own experience has been to just read through them quickly, and make fast decisions, so I don't want to do too much, but I am working without a scale. So I don't want to do too much revising, but I am reconsidering, but I am going to look at them again just to see whether my ... the scores are consistent or not. Uh, I am concerned about that and recognize that I only gave one 6, but that was distinctly more effective writing than any of the other ones. Um, I gave a lot of, I think I gave, I think I gave a lot of 1s, 2s, and 3s, particularly 3s, and that seemed a little off balance but that just may be where the sample is, what was selected for me. (Roy)

These macrostrategies had a procedural aspect: the raters used them to guide themselves overall through the relatively undefined task of reaching assessment judgments. They also had a substantive aspect: the participants established specific criteria on which they based their judgments, based on their consideration of the unique set of compositions each person received. Similarly, most participants expressed an awareness of their consideration of specific aspects of the written compositions, based on their previous evaluation experiences:

> Again, I don't know why but it seems that I am working like a machine, looking at content and organization first and language afterwards. And once again, all I can say is that for the past 3 or 4 years I've basically been working with the English Composition Ability Profile, based on Jacobs et al. (1981), where the, uh, thing is broken down essentially into language, organization, content, vocabulary, language, and mechanics. I tend to conflate vocabulary into language, uh, and I tend to look at it more holistically than they do, but I guess in, in a sense what I'm doing is separating organization and content, uh, and, uh, in considering them first and then looking at the language, and I guess I must be equating the two somehow or averaging them out. (Scott)

Interestingly, as they established their criteria for particular score points, most participants noted that they had to envision the abilities and personal situations of the examinees to make a decision on scoring — for instance, within a university context:

> It's hard for me to know whether I'm interrelating these to just the whole set of compositions themselves. And I'm still hoping there are some better pieces of writing that I'll give a 6 to. Or to give some criteria like, is this person ready for university admission? I suppose this person could do an undergraduate degree but they probably would not get very high marks in, uh, their writing. But they seem to have a grasp of English that would enable them to do so. So, I'm waffling between whether I'm doing a relative kind of rating here and in respect to the set of compositions that I have. Or whether I'm, I'm trying to set them against some criteria like readiness for university. (Roy)

Such behaviors establishing criteria for the assessments, and indeed any of these macrostrategies, may not have occurred if the assessment task had been more specifically prescribed for the raters, such as would probably occur in the administration of a composition test. Raters of composition tests are typically provided relevant criteria for assessment in advance, along with sample or anchor papers for specific score points. Here, the raters had to identify these criteria themselves, as Zoey, for instance, spent considerable time conscientiously doing. In other words, these macrostrategies may in fact be artifacts of the way we defined (or rather did not define) the assessment task, leaving raters to make definitions of scoring criteria themselves. This in turn may have led to behaviors like sorting the compositions, rather than adhering to a specific rating scale prescribed in advance.

Sequences of Decision Making While Rating: Prototypical Sequences

Nonetheless, nearly all of the decision-making behaviors in the think-aloud data involved micro-level strategies — particular decisions focused on assessing a single composition, such as the majority of those listed in Table 1. In reviewing the transcribed data, we identified the following sequence as prototypical of the participants' assessment behaviors overall:

1. Scan the composition for surface-level identification, such as length, format, paragraphing, script (e.g., typed or handwritten).

2. Engage in interpretation strategies, reading the essay while exerting certain judgment strategies, such as:

    - classifying error types (lexis, syntax, morphology, spelling), leading to an assessment about the command of language

    - identifying comprehensibility, leading to an assessment of language use and rhetorical strategies

    - interpreting rhetorical strategies (in terms of relevance, rhetorical knowledge and performance, coherence, redundancies, topic development), leading to an assessment of content and organization

    - envisioning the situation and personal viewpoint of the writer

3. Articulate a scoring decision, while summarizing and reinterpreting judgments.

A prototypical example of this sequence is:

Next essay is essay 23, about the room. I haven't had one of those in a while. It's typed. I find typed ones are much easier to look at, deal with. It's three paragraphs. Uh ... uh ... O.K., there's, the first short paragraph is basically which room is the most important and why, a short explanation of why. The second paragraph's going into more detail. (Reading). O.K., there seems to be good use of sentences, good use of vocabulary. There's a good use of, haven't seen this in a lot of the essays, painting a picture. For example, "WE CAN EVEN ..." It's a long sentence, but it's cohesive, it paints a picture, it deals with different issues. And the last paragraph talks about the other rooms in the house, that they are useful. Why they are useful ... but then goes back to the bedroom, and says if not for the bedroom. Uh, ... that's quite nicely written, actually, "BUT IF WE ..." this is very nice. "NO, WE DEFINITELY ..." I would definitely give this a 6. I, uh, don't see

> a lot of surface errors, structures, on first glance. It's very well organized, very well structured. Uh, the reader's brought into it as well. Uh, as I said, a very good picture is painted for the reader, uh, of the writer's opinion, why it is so. There's good use of adjectives as well. Uh, it's very well structured, it's very coherent. Certainly all very relevant. And, so the overall organization, the overall language use is good. Definitely I'm going to give this one a 6. Which is neat. We're able to see the differences between a well written, you know even if it's, this isn't an incredibly long, it's three short paragraphs, but it's very well done. (Karen)

As this prototypical sequence and Karen's example suggest, participants integrated their decision making in complex, interactive episodes of information gathering, reasoning, and hypothesizing, prior to making, then evaluating, their assessment judgments. The decision-making behaviors presented in Table 1 appeared in the verbal reports not so much as fixed, isolated behaviors, but rather were interdependent sequences of attention to variable facets of the compositions and the criteria the raters considered relevant to judging them. Interpretations derived while reading, and responses to features of the written compositions, were formulated gradually and interactively, balancing one kind of impression against others (e.g., overall length against originality of rhetoric, as in Karen's example above). They developed into judgments of quality, then scoring decisions, for each essay. Raters arrived at some assessments quite quickly, within a minute or so, whereas other decisions extended over one or two pages of transcribed verbalizations.

But Karen's example, immediately above, can only be termed *proto*typical, because the raters in fact displayed — both in contrast to one another and within individual think-aloud protocols — several variants on this sequence. For instance, some raters tended to focus their initial interpretation strategies on identifying ideas and propositions, then classifying rhetorical strategies, then classifying error types, before finally reaching and summarizing their scoring decisions:

> O.K., next one (coughs), excuse me, is 130. 130 is the topic of vacation. Uh, let's see. This is typed. It's fairly long, there's four paragraphs here. "PEOPLE ALWAYS NEED ..." O.K., so in the first paragraph they are basically setting up the arguments for the two types of vacations, and she closes the first paragraph, or he, the writer closes the first paragraph with their opinion. So that's very well organized. Uh, then the next two paragraphs, one paragraph discusses the concept of long vacations, uh, and gives examples, personal examples and examples of other people she knows or he knows ... She's more personal, which is good, it's backed up, everything she says is backed up with examples. Uh ... "AFTER A LONG ..." Then the third paragraph talks about short vacations ... the recharging, "IT RECHARGES ME." It's a good expression. O.K., then the last paragraph again reiterates her opinion, or his opinion. "HAVING SHORT VACATIONS ..."

> O.K., well, basically the structure's good. There's an introduction, there's a conclusion, there is a main body, which is quite well organized. There's a structure to the argument. The arguments are all backed up with examples. Uh, there are some problems in terms of ... I'm trying to think of what is it exactly. There's something that's a problem in terms of the sentences, the structures. Uh, ... there are some problems with prepositions and use of those. Uh ... there's good use of linking words. The sentences are linked up quite well, the ideas are linked quite well ... Problems with, uh, plural and singular in terms of the nouns, when she's talking about "students" using "student." I see that several times throughout. In terms of school, with vacation ... challenges, pressures, hm. Bit of trouble linking the verbs ... to who they're referring to. Agreement. Basically it's pretty good though. Uh, mmm, trying to decide between a 4 and a 5. It becomes more difficult the more you see, becomes easier, and it becomes more difficult. I'm giving this one a 5. I think basically the sentence structure is good, sentences are complex, they're not simple. The writer's attacking I think the idea in a very good, logical, coherent way. It's relevant, it's interesting to read. The topic is certainly developed. Uh, ... O.K. And there's, there's, certain, you can see certain problems, errors along the way, but pretty much I think it's a pretty decent paper. O.K., I'm going to give that one a 4 or 5. Five, I'd say, I'm giving it a 5. (Karen)

In contrast, other participants interspersed their attention to language aspects with their attention to rhetorical aspects. For example:

> Next essay, 117, and I'll be reading that ... Second sentence ends strangely and abruptly. That is a bad sentence to me. I'm reading on ... Uh, here in the development of the first argument the author is sliding off topic. There can be some connection found, but not a good one ... O.K., the second argument, that sentence is just wrong, bad sentence structure, bad word choice. I don't even quite understand the meaning of all this. I'm reading on. Oh, he or she is done by the end of the third line, uh, with the, uh, discussion of the second argument. I'm reading on. I think I can guess what the author is trying to say but it is not a clear thought ... O.K., end of essay 117. Obviously this person knows about framing, uh, there is an introduction, there is a one sentence conclusion, uh, and three arguments, of which only the first one is close to the actual argument development that, uh, I would like to see in an essay. Oh, it's either a 3 or a 4. Well, the language is not good, many errors, structure, sentence structure, errors, and word choices, grammar. "THERE IS NO ..." and so on. That's a 3, 117 gets a 3. (Patricia)

And some sequences of decision making focused largely on identifying language errors, particularly for compositions raters scored low. For instance:

> Essay number 107. I just finished 74, and this is essay 107. This is only one paragraph, and it's typed. I'm going to read now ... O.K., I finished reading. This is essay 107. Uh, it's only one paragraph. Sentence grammar is very poor. Punctuation is missing. The student knows how to use periods, but not capitals. The student does not use any commas, and there are also many spelling errors, an incredible amount, and basic spelling errors for fairly simple words, such as "I DON'T KNOW," D-O-N comma N. Looks like this student had trouble typing. So that's my guess because of this error. "I DON'T KNOW ... " It looks like "i" should have been "it." There's a missing "t" there. The final sentence is "I AM SO ..." The student is apologetic for turning in such a poor essay. I give this a 1 on the scale of 6. I move on now. That's the end of essay 107. (Paul)

Text Features Attended to While Rating Compositions

To address subquestion 3 — *Which text features of the essays do the instructors/assessors attend to for assessment purposes, as documented in concurrent verbal reports of their thinking while assessing them?* — we prepared a comprehensive tally of text features ("traits," as Hamp-Lyons, 1991, calls them, or "rater objects," as DeRemer, 1998, refers to them). We culled these from the entire set of think-aloud protocols, listing all text features that raters attended to with any frequency (i.e., more than once). The focus of this analysis was on aspects of the compositions to which raters reported attending, rather than qualities of the decision-making behaviors they demonstrated, though there is an obvious interaction and overlap between the two. The resulting list of text features is presented in Table 2.

This approach to analysis, if extended in future research, may be useful in developing scoring rubrics — along the lines of primary trait scoring for particular types of writing tasks (cf. Lloyd-Jones, 1977) — for the TOEFL 2000 examination. As Vaughn (1991, pp. 122-125) suggested in proposing a similar list from her analysis of think-aloud protocols from five raters of ESL compositions, such text features could be analyzed for raters' expressions of either positive or negative impressions (though certain items, such as those listed in Table 2 under Style, would warrant rephrasing to accommodate that orientation). However, prior to such analyses, the present list of text features must be refined and evaluated further, and various problems must be resolved, including problems of:

- terminology (i.e., the use of differing terms to refer to similar phenomena, or alternatively, common terms to refer to differing phenomena)

- categorization (i.e., distinguishing comments at the global or local level of texts; determining interrelations between different aspects of texts, such as rhetoric, language, or vocabulary)
- orientation to theories of language or grammar
- the extent or appropriate length of the list
- reliability

Such limitations may defy validation of such a list of text features unless a particular theoretical orientation to text analysis is adopted. At present, the terms in this list seem primarily to represent concepts common to ESL/EFL instructors' discussions of student texts (i.e., a kind of checklist of basic things that teachers conventionally state about writing English). The list is derived empirically from the think-aloud data, rather than a particular theoretical conceptualization of what written English is or should be. We have not attempted to link the categories in Table 1 directly to those in Table 2 because the two derive from different analyses, though there are obvious areas of overlap. The categories in Table 1 are more robust, encompassing most of the higher-order categories in Table 2. See further discussion of these matters in our account of Phase Two of this study.

Table 2

Text Features to Which ESL/EFL Instructors/Assessors Attended While Rating TOEFL Essays

Layout	Rhetorical organization	Style	Topic Development
a. Handwriting b. Typing c. Spacing d. Lineation and paragraphing e. Legibility f. Editing or self-corrections g. Upper/lower case script	a. Essay structure ☐ Introduction ☐ Title ☐ Topic orientation ☐ Thesis statement ☐ Sequencing ☐ Body ☐ Conclusion b. Paragraphing ☐ Number of paragraphs ☐ Development of paragraphs ☐ Format of paragraphs ☐ Topic sentence c. Cohesion d. Narration e. Dialogue f. First person g. Second person h. Third person i. Reader involvement	a. Clarity b. Colloquialness c. Communicativeness d. Comprehensibility e. Formulaic f. Sophistication g. Idiomaticity h. Literalness i. Interestingness j. Conciseness k. Specificity l. Strangeness m. Convincing n. Effectiveness o. Exaggeration p. Cliches q. Repetition r. Register s. Fluency t. Literacy u. Creativity v. Ambiguity w. Awkwardness x. Presumptuousness y. Triviality z. Moralistic aa. Consistency bb. Avoidance strategies	a. Argument ☐ Logic ☐ Reasoning ☐ Development ☐ Examples ☐ Analogies or metaphors ☐ Supporting evidence b. Task completion c. Topical focus d. Relevance e. Coherence f. Persuasiveness g. Misunderstanding
Total written production a. Length of essay b. Length of paragraph c. Length of sentence		**Sentences** a. Sentence structure b. Sentence complexity c. Number of sentences d. Itemized lists e. Sentence variety f. Completeness of sentences g. Run on h. Questions	**Grammar** a. Preposition b. Pronoun c. Subject-verb agreement d. Tense e. Auxiliary f. Article g. Number (singular/plural) h. Adverbial i. Adjective j. Relative clause k. Adjective clause l. Gerund m. Participle n. Verb o. Subject p. Object q. Passive/active voice r. Noun phrase s. Conjunction t. Conditional u. Comparatives and superlatives
Vocabulary a. Choice b. Range c. Variety d. Precision e. Morphology f. Idioms g. Missing words			
Punctuation a. Age b. Gender c. Ethnicity d. Mother tongue e. Location f. Beliefs g. Motivation h. Prior education or writing experience i. Potential or readiness for learning or university studies	**Personal situation of writer** a. Comma b. Period c. Capitalization d. Question mark e. Apostrophe f. Quotation marks g. Colon h. Semicolon i. Hyphen	**Errors** a. Quantity b. Gravity c. Frequency d. Typographical errors	
		Spelling	

Variations in Raters' Decision Making

The fourth subquestion — *Do these raters' decisions and the text features they perceive vary appreciably with the examinee's level of English proficiency (as determined by previous TOEFL ratings), with the essay topic, with particular reader characteristics, or with individual raters?* — was difficult to answer concretely before we undertook a systematic coding of the think-aloud protocols, which we did in Phase Two (see later sections of this report). Nonetheless, for Phase One, we reviewed the think-aloud data to identify obvious qualitative trends in the data, then held subsequent discussions among the research team to consolidate our impressions in respect to subquestion 4.

First, we noticed that raters appeared to attend to certain types of text features based on score level — though, as noted above, the trend was not a consistent one. For instance, on essays the raters scored low (e.g., as 1 or 2), they attended mainly to surface language features of the compositions. For example:

> The next essay is number 96. It is typed ... and appears to be in three paragraphs. I'm just taking a look at the topic sentence, "THERE IS NO ..." ... All right. "STUDYING DIFFERENT AREA ..." Spelling mistake there, "THE STUDENT VERBAL ..." Spelling mistake. "SKILL. I AM ..." O.K., so it's sounding very nonidiomatic, uh, incoherent, major structural problems here. "IMAGINE THIS, ONE ..." O.K., not logical, serious grammar problems. "HOW DO YOU ..." Already I, the use of questions tends to irritate me anyway, but, they're, it seems to me that they don't have anything else to say, so they're writing a list of questions. "CAN I USE ..." Uh, O.K., structural problems there, a few typing or spelling mistakes, too. "SOME PEOPLE STILL ..." Spelling mistake. "THAT P-O-Z-I-T-I-V-E." Spelling mistake. "SCIENCES IS IMPORTANT ..." Illogical, incoherent sentence there. "STUDYING DIFFERENT SUBJECT ..." Uh, although they have attempted to write and address the topic, they have not at all, they haven't even answered the question, mmm, fence-sitting, they haven't taken a stand. Incoherent. Uh, ... , I think I'd give this one a ... 1. (Jane)

At the opposite extreme, raters seemed to consider mainly the rhetoric and ideas in essays they scored high (e.g., as 5 or 6), often engaging verbally with the author's argument or position. For instance:

> Number 18 ... The room in the house one. This is the longest response that I've seen so far, a full typed page to the room in the house. One, two, three, four, five, six, seven paragraphs. The first and last consist of only one sentence. The first is a topic sentence: "FOR A YOUNG ..." O.K., now, the person starts talking about

him- or herself. “I AM A ...” So he starts telling a story ... So the person’s made it totally personal. They’ve talked about their own house now, and then explained why it’s difficult to pick a room, and they are going on to say why the dining-living room ... Just reading ... Jumps around a little bit. “THE LIVING ROOM ...” Now, he’s describing how their room is laid out. “SINCE THE TV ...” O.K., now, this is interesting, because they are getting to the point. They’ve written one, two, three, four, five paragraphs on their own personal experience, and now they’ve decided to address the question, which is not “Which is the most important room in your house,” but which is the most important room in a house, I think. Yeah, “THE MOST IMPORTANT ...” Now they get to it. “LASTLY, I WOULD ...” Uh, I’m torn. I’m really stuck with this one, because I like some of the, I like some of the language. “I AM A ...” So, it’s very well written. Very well written. And yet, I wouldn’t have organized it, an essay, in this way. And I guess maybe I’m kind of stuck on the, the traditional give the general information first. So, in other words, I would have put the paragraph about, you know, “LASTLY, I WOULD ...” I would put that first. That the most important room depends sort of on who you are, and what your lifestyle is. Then, maybe it would have been O.K. to go on and have the personal experience, and then conclude. So that’s an organizational preference, perhaps, rather than a, than a general rule. And since I can’t fault the writing, I don’t think that I can mark this down from a 6 to a 5 because of a personal preference. So, I’m going to give it a 6, because I think it’s a well written piece. (Zoey)

In the middle range of writing proficiency (i.e., essays scored 3 or 4), raters’ decision making consisted largely of balancing positive and negative factors related to content, rhetoric, and language. They often expressed some uncertainty or ambivalence as to how to judge or equate these elements. For example:

O.K., the next one is 128. Another composition about vacation. O.K., typewritten. One, two, three, four, five paragraphs. Um ... looks long. “BEING A HIGH ...” “Variant?” Is “variant” used in the proper way here? I don’t know. But nice introduction. My, although slow. It is a bit slow. Nice introduction. “IN MY OPINION ...” The use of “also” here is a bit weird. Um, how does he calculate? So precise. Giving the number of weeks. Interesting ... Nice ad, use of adverbial phrases. “I THINK THAT ...” A bit colloquial here. “TOO LONG THOUGH ...” This combination seems a bit strange to me. “THEY SHOULD BE ...” His comment ... His personal comment, opinion. Nice. “I DISAGREE WITH ...” Why? “IN MY OPINION ...” Oh my, use of words is neat, “prolonged studying” and “getting too much information,” blah, blah, blah, “hence.” “ALSO ONE LARGE ...” “Disruptive”? “Disruptive”? What’s that word? “RESULTING IN LOWER ...” O.K. “FOR EXAMPLE IN ...” You’re talking about the present. You

> say “take.” “USUALLY TAKE FROM ...” “Find” or “found?” No, the order is wrong. “FIND IT VERY DIFFICULT ...” No, no. Until here it sounded so good. O.K., “SO FROM WHAT ...” I don’t understand this conclusion. This conclusion doesn’t fit properly in the structure, in the whole style of presentation. A bit disappointing. But ... yeah ... has the idea of presentation. This presentation idea is pretty clear. There’s an introduction. Tried to put a conclusion which does not really fit in the style, and the body is ... Gave examples. Un, nice. Some awkward phrases, like use of “also” at the end of the sentence. And then grammatical, grammatically awkward sentences. Um, one is “me and my friends,” “found very it difficult.” This word order and location is not proper. Not proper. Level 3. Do I give it a level 4? Level 3. (Melissa)

Second, we could not discern any differences in raters’ decision making across the four different essay prompts or topics (which bodes well for the design of the TOEFL essay prompts), though we did not carry out any numerical analyses of this in our research (for the reasons noted above). Several raters, however, observed that the topic related to a university seemed more complex than the other three, demanding a more extensive argument from examinees. Nonetheless, this did not appear to affect the assessment behaviors documented in the think-aloud protocols.

Third, we observed certain differences in individual styles of rating, though as noted earlier, raters’ styles were somewhat uniform overall. As discussed earlier, most Phase One participants with extensive experience assessing ESL/EFL compositions tended to verbalize their decision making more extensively, producing longer, more detailed think-aloud protocols than less experienced assessors. Also, some of the more experienced assessors adopted particular macrostrategies — such as sorting the essays into piles and reviewing their initial decisions collectively at the end of the research task — to guide their judgments and confirm their scoring criteria. Finally, some of the raters were prone to focus more on reading and interpretation strategies (e.g., Gary), whereas others focused more on judgment strategies (e.g., Jane, Paul).

4. Phase Two: Revising the Framework

Introduction

In Phase Two we refined and evaluated certain preliminary findings from Phase One. In addition, we collected and analyzed further verbal report data generated while a) expert raters of native-English compositions assessed TOEFL essays and b) ESL/EFL instructors/assessors who participated in Phase One rated five TOEFL 2000 prototype tasks. Having developed and piloted our instrumentation, methods of data collection, and preliminary analytic framework in Phase One, our intention for Phase Two was to concentrate on coding and analyzing the existing and new verbal report data to verify our initial findings and to evaluate the prototype tasks. As in Phase One, in Phase Two we pursued one major research question and four related subquestions, as follows:

Does the analytic framework developed in Phase One account for the assessment behaviors of the same instructors/assessors on a different sample of written compositions (i.e., five TOEFL 2000 prototype tasks and other TOEFL essays) and for the assessment behaviors of another sample of highly experienced and highly regarded raters of native-English composition? What further refinements, developments, and research are needed on the descriptive model and in the prototype tasks?

1. To what extent is the analytic framework developed in Phase One able to account for the self-monitoring and decision-making behaviors of the same assessors/instructors who participated in Phase One on a different purposive sample of TOEFL essays — five prototype writing tasks produced by another TOEFL 2000 project (Santos & Kantor, n.d.)?

2. To what extent is the analytic framework developed in Phase One able to account for the self-monitoring and decision-making behaviors of a sample of highly skilled and regarded raters of native-English composition while they assess a purposive sample of TOEFL essays?

3. What refinements, developments, and further research are needed for the descriptive model of raters' decision making?

4. What refinements, developments, and further research are needed on the TOEFL 2000 prototype writing tasks?

We felt we had such extensive representation in Phase One from raters with ESL/EFL experience in Canada and internationally that it would be worth checking our initial findings — particularly for the TOEFL essays — against an independent group of raters with extensive native-English-composition assessment experience in the United States. Also, since we collected more data in Phase One from the raters in Toronto than we had initially proposed — and since

we received data on five TOEFL 2000 prototype tasks, rather than the two tasks we had initially expected — we decided to abandon subquestions 1 and 2. On the basis of the findings and the extent of the data initially collected for Phase One, these now seemed unnecessary.

Method

Seven of the 10 experienced instructors of ESL/EFL in Toronto who participated in Phase One (with one substitution in personnel), plus seven highly regarded and experienced assessors of native-English compositions from various locations in the United States, participated in Phase Two. The latter were nominated by ETS staff who organize scoring of the writing components of the SAT II: Subject Tests, the GRE exam, and other College Board® and ETS assessments. None of the U.S. raters had previously had any contact with the research team in Toronto, so they could not be expected to be influenced by Phase One of this project. To document their decision-making behaviors, all participants produced concurrent think-aloud protocols while rating, as in Phase One; they recorded these on audio cassettes following preset instructions (see Appendix B). In addition, they later completed a profile questionnaire (see Appendix A).

Toronto raters. For Phase Two, the seven raters in Toronto each assessed a total of 36 compositions. Six adult ESL students at a university in California that was participating in a concurrent ETS project (Santos & Kantor, n.d.) each responded to five TOEFL 2000 prototype tasks plus one TOEFL prompt. The five prototype tasks involved these ESL students in:

- Task A, writing a summary of a lecture (on voice or geology) they had heard
- Task B, writing a note or a summary after listening to a conversation (about teaching assistants or an election)
- Task C, writing a summary of a reading passage (on beads or urbanization)
- Task D, writing a note or summary after reading a conversation (about a landlord)
- Task E, writing a response after reading a lecture (on urbanization or diet)

The 36 compositions were presented in a unique random sequence for each ESL/EFL rater, but they were grouped by author so that the raters would not have to guess which compositions were written by which of the ESL students (which we expected they would inevitably do with this number of compositions). As in Phase One, these seven raters collaborated as both participants and researchers on the design, data collection, analyses, and interpretation of findings. They attended weekly meetings from September through mid-December, working in pairs on specific aspects of the research. All seven participants generated

think-aloud protocols while rating the set of 30 compositions written for the prototype tasks and the six TOEFL essays (a total of 252 protocols), they transcribed another person's protocols, and they later verified the transcriptions of the think-aloud data they had themselves produced.

U.S. raters. ETS staff, at our request, initially identified 12 potential raters in the United States who would be especially suitable for this research, based on their extensive experience assessing native-English compositions for various tests administered by ETS. They were initially contacted in early October by email from Alister Cumming (to maintain confidentiality of their participation at ETS). They were invited to participate in the study for a stipend of $200 and were asked to sign consent and confidentiality forms. The first seven who responded were selected. Four others declined because of other work commitments, and one who said she was willing to participate had replied after the quota of seven had been filled.

These raters were located in different cities in the northeastern or southern United States, and were typically employed as professors, instructors, or program administrators at a university or college. Each received a package of 40 TOEFL essays, randomly selected from the same pool of 143 essays as in Phase One (including four different essay prompts), and sequenced in a unique random order for each person, along with samples of the essay prompts, several cassette tapes, instructions for rating the compositions and producing think-aloud protocols (as shown in Appendix B, with minor modifications to suit the context), the profile questionnaire, and a cover letter itemizing the contents of the package and directing the return of the package. When the completed packages were returned a few weeks later, the tape-recorded data were transcribed in full by the Toronto team.

Participants

Toronto raters. Characteristics of the seven-member Toronto team — graduate students enrolled at University of Toronto's Ontario Institute for Studies in Education — are described earlier in section 3 of this paper. Since Phase One, two students completed their master's degrees and one accepted other research employment for the term, so they did not participate in Phase Two. One new rater with a similar background to the other graduate students joined the team in September. He was in his late 20s; had spent several years teaching English in Canada and overseas; had some previous research, editing, and assessment experience; was a native-English speaker; and was presently enrolled in a Ph.D. program in Second Language Education. To preserve confidentiality, the pseudonym for this person is included among the others listed for Phase One.

U.S. raters. We gave the seven U.S. raters the pseudonyms Charles, Doug, Fred, Kathy, Jean, Lucy, and Sam. According to information they provided by way of the profile questionnaire (see Appendix A), the U.S. raters had the following characteristics:

- They were in their 30s (one person), 40s (three people), or 50s (three people).
- Four were male and three were female.
- Each had exceptionally extensive (from 10 to 41 years) experience assessing and teaching English composition to adults in academic contexts in the United States.
- Each possessed either a Ph.D. in English literature (five people) or a related field (one person), or a master's degree in English writing (one person). Only one person possessed a teaching certificate, though most had received specialized rater training through ETS and/or their university.
- Each currently worked a) as instructor/assessor of English writing (four people), b) as an instructor of English (two people), or c) as an administrator of an English writing program (one person).
- All said they had worked primarily in native-English writing for adults, but two people also had some experience with ESL assessment and instruction.
- All were native speakers of English, their dominant language (though one grew up in a bilingual family).
- All rated their abilities to assess English writing as "expert," except for one person who rated herself as "competent" (based on only three years' assessment experience, but 20 years' experience teaching English composition).
- All had extensive experience writing professionally — either a) as authors of numerous books and articles and editors of scholarly publications (five people, two of whom also had translation experience) or b) as writers and editors of technical, journalistic, business, or professional publications (two people).

Approach to Analyses

Our analyses in Phase Two focused on a) coding the think-aloud data with categories derived from the preliminary descriptive framework developed in Phase One, b) refining the descriptive framework, and c) applying it to samples of think-aloud data. We wanted to revise the framework and to evaluate how it fared in accounting for data generated by U.S. raters for the TOEFL essays and by Toronto raters for the TOEFL 2000 prototype tasks. In addition, we wanted to use the coded data to assess empirically certain impressions that we had from reviewing qualities of the Phase One data, specifically:

- What was the frequency and distribution of each of the decision-making behaviors that raters used when assessing TOEFL essays?
- Did the TOEFL 2000 prototype tasks produce similar frequencies and distributions of decision-making behaviors?
- Were there differences in the extent of raters' attention to language or to rhetoric and ideas in TOEFL essays scored low or high?
- Were there differences in the extent of raters' attention to language or to rhetoric and ideas between the ESL/EFL instructors/assessors in Toronto and the native-English-language instructors/assessors in the United States?

Answering these questions empirically would also address certain of the research questions guiding Phase Two.

To begin this process, all members of the research team in Toronto worked together to segment and code initial samples of think-aloud data, to discuss points of agreement and differing interpretations, and to develop guidelines for coding, as well as to make minor revisions to the framework based on these initial trials. In doing the latter, our criterion was whether the Phase One framework could be applied reliably to the data by different members of the research team. We revised or rejected categories that could not be. In addition, we selected exemplar samples of statements from Phase One think-aloud data to anchor our coding of each category; these are displayed in Appendix E. The revised version of the framework, which we used for coding Phase Two data, appears as Table 3 later in this report (see *Findings*), and features several minor changes from Table 1.

We used three criteria to segment the think-aloud protocols into separate, comparable units of decision making. Pauses of five seconds or more (indicated by ellipses in our transcriptions), the reading of a segment of the composition by the rater (marked by capital letters in quotes in our transcriptions), and the start or end of an assessment signaled a new segment. These criteria are similar to those suggested by Ericsson and Simon (1984) and used by Cumming (1990) for analyses of concurrent verbal reports of thinking processes.

In this manner, we produced a summary count of segments for each think-aloud protocol for each composition rated (ranging from about three to 65 segments per composition per protocol), as well as a tally of the categories of decision making coded for each segment. We considered each segment of decision making to involve at least one, but potentially several, decision-making behaviors, which most did. Two members of the research team then worked to establish an average intercoder agreement of 84% on 25 independently coded think-aloud protocols produced by all 10 raters for the four Phase One TOEFL essays (agreement per

protocol on 587 coding decisions ranged from 73% to 100%). Next, each of these two coders worked independently to code the think-aloud data from both Phases One and Two, then worked together again to resolve discrepancies between themselves before submitting the coding for analyses. Later, in checking their agreement on independent coding of all coded data from the native-English-composition raters, the two coders demonstrated 90% agreement overall on 358 coding decisions. The data were entered into Microsoft® Excel spreadsheets by four members of the research team, who checked each others' entries. We then conducted analyses using the data analysis software product SPSS®.

We selected only a subset of the think-aloud protocols for the TOEFL essays for coding. As noted earlier, because the TOEFL essays that each participant received were randomly selected from a larger pool of 143 essays, and uniquely so for each person, there was not a common set of compositions identified in advance for comparisons. Moreover, since we had not defined a common rating scale for the raters (because we wanted instead to see how participants put their own interpretations on their ratings), the scores that people gave to each essay tended to vary. Thus, we selected for coding only those protocols representing essays scored consistently by most raters, and we required that the essay protocols selected reflected the full range of score points (i.e., from 1 to 6).

We reviewed the Phase One data to identify TOEFL essays scored consistently by all (or almost all) raters. We found 106 protocols that met our criteria, including:

- seven essays consistently rated as **1** (the lowest score point) in 39 protocols
- four essays consistently rated as **2** in 27 protocols
- three essays consistently rated as **3** in 17 protocols
- one essay consistently rated as **4** in 6 protocols
- three essays consistently rated as **5** in 12 protocols
- two essays consistently rated as **6** (the highest score point) in 5 protocols

We then used these 106 think-aloud protocols, generated for 20 essays scored consistently by Phase One raters, as the basis for our remaining analyses. We felt we could treat these protocols as comparable for the purposes of our descriptive statistical analyses. But, as noted earlier, the quantity and distribution of these protocols, and of the compositions on which they were based, were too few (from the total of 143 essays) and too skewed (toward the lower end of the scale of ratings) to be amenable to analyses with inferential statistics.

To facilitate pair-wise comparisons of Phase Two data, we selected and coded all of the think-aloud protocols produced by the seven U.S. native-English-composition raters for the same 20 TOEFL essays listed above. We found 31 matches, then we used a "lottery method" of random selection to pair these protocols with comparable Phase One protocols. The two participants who had established high intercoder reliability on the think-aloud protocols for the TOEFL essays independently coded all of the Phase Two think-aloud data generated by Toronto raters for five prototype tasks and one TOEFL essay produced by each of six ESL students. Some of the compositions written in pencil by the ESL students on the TOEFL 2000 prototype tasks did not photocopy legibly, and as a result some raters were not able to read or score them, creating 42 instances of missing protocol data. However, these perceptions and decisions varied among individual raters, so at least some think-aloud protocols and ratings were generated for all of the prototype tasks. In total, we coded 210 think-aloud protocols for this group of essays. Unusually for this type of research, we did not experience any problems with missing data due to technical problems with any of the tape recordings.

To answer the four questions posed at the beginning of this section, we converted all of the coded data to percentages of the total number of decisions made by individual raters for each composition. Converting the frequencies to percentages was necessary because, as is typical in think-aloud data using large numbers of coding categories, raw counts of decision making or segments varied greatly from rater to rater, and so were not amenable to comparative analyses in that form. Also, because each of the 35 coded categories of decision making proved to account for relatively small portions of each raters' overall decision making (see Tables 3, 5 and 8 in the next section of this report), for most analyses we grouped the data under larger categories reflecting those presented in Table 1 (i.e., language focus, rhetorical and ideational focus, etc.). We then used means of percentages of decision-making behaviors for groups of raters (i.e., ESL/EFL instructors/assessors or native-English-language instructors/assessors), for composition tasks (i.e., TOEFL essays or prototype tasks), and for essays with similar scores (i.e., those rated as high or low) to compare similarities and differences among them. That is, the units of analyses in the statistics reported in the section that follows are segments of decision making that occur in the think-aloud protocols — not the rater, the essay, or the essay topic. These are not independent observations; the same rater may have generated several of the decision-making behaviors for one or more of the TOEFL essays.

We also reviewed the TOEFL 2000 prototype tasks. Two members of the research team interviewed the participants in Toronto (including each other) to identify strengths and weaknesses they perceived in these new tasks. The purpose of this analysis was primarily formative, aiming to provide recommendations to improve the design of these writing tasks from the viewpoint of experienced ESL/EFL assessors in the future.

Findings

Following the research questions guiding Phase Two, as well as issues arising from Phase One results, we present findings for Phase Two of the study in respect to the following:

- characteristics of participants influencing their ratings
- decision making by native-English-composition raters
- decision making for TOEFL 2000 prototype tasks
- refinements to the descriptive framework of decision-making behaviors
- statistical analyses of coded think-aloud data
- formative assessment of TOEFL 2000 prototype tasks

Recommendations based on these findings, together with findings from Phase One, appear as the final section of this report.

Characteristics of Participants Influencing Their Ratings

Native-English-composition raters. An important issue arising from Phase One was how raters perceived their scoring behaviors to have been influenced by their previous experiences. In Phase Two, the native-English-composition raters in the United States indicated (in answering section I, item 1, of the profile questionnaire) that they thought their assessment behaviors were influenced by particular assessment schemes they had used previously "a great deal" (self-ratings of 5 by two raters and of 4 by four raters) or "slightly" (self-rating of 3 by the one rater with relatively less assessment experience). In describing these influences (section I, items 2 and 3 of the questionnaire), they all cited their experiences with a variety of tests administered by ETS — such as the SAT II: Writing test, Advanced Placement Program® examinations, the Graduate Management Admission Test®, the GRE Writing Assessment, The Praxis Series™ assessments, the TOEFL and TWE exams, and College-Level Examination Program® tests — as major influences. Other major influences included their own teaching and writing assessment experiences and work with students at universities, and one rater mentioned a published research study he had read. Given the raters' extensive experience as composition assessors (which was our criterion for selecting them), and our decision not to prescribe a rating scale for participants, this high degree of influence from other scales for rating writing was probably to be expected.

These native-English-composition raters described the "three qualities [they think] make for especially effective writing in the context of a composition exam" (section I, item 4, of the profile questionnaire) in relatively uniform terms. These qualities can be summarized collectively as follows:

- rhetorical development, argumentation and organization of ideas, achievement of purpose and focus, logical reasoning, and uses of evidence
- coherence and clarity as well as accuracy, variety, and fluency in English

A few raters also mentioned unique terms in this part of the questionnaire, such as "stylistic flair," "audience awareness" or "concerns for the readers' understanding and continued interest," and expression of a unique "authorial voice." All but two of the U.S. raters mentioned accuracy of English language in their responses to this questionnaire item, but they did not seem to put the same premium on this aspect of writing as the ESL/EFL assessors in Toronto did. As observed earlier in discussing parallel findings from Phase One, the terms people may use to describe their criteria for evaluating writing vary and are open to variable interpretations.

ESL/EFL composition raters. After rating the TOEFL 2000 prototype tasks in Phase Two, the seven ESL/EFL instructors/assessors indicated that they were influenced by previous experiences (section I, item 1, of the profile questionnaire) as follows:

- "a great deal" (self-rating of 5 by one rater who cited involvement in Phase One of this research as influencing him a lot; self-rating of 4 by one rater who described as an obvious influence a particular rating scheme he uses regularly in his teaching)
- "slightly" (self-ratings of 3 or 2 by three people who also said they were influenced by Phase One of the research)
- "not at all" (self-rating of 1 by the one person who had not participated in Phase One)

Descriptions provided by ESL/EFL instructors/assessors of the "three qualities [they think] make for especially effective writing in the context of a composition exam" (section I, item 4, of the questionnaire) were similar to the responses they gave in Phase One of the research (reported earlier). These descriptions were more variable from rater to rater than those provided by the native-English-composition raters. Also, the ESL/EFL instructors/assessors emphasized formal text characteristics and particular language and writing abilities more extensively than the native-English-language instructors. It is difficult to know how to interpret these differences between groups of raters (or between raters' Phase One and Phase Two self-assessments), other than to suggest that there may be many possible influences — such as, a) differences in individual work experiences, b) differences in predominant professional conceptualizations

common to the fields of either ESL/EFL instruction or native-English-composition instruction, c) effects of experience with holistic scoring among the U.S. raters, d) influences from participating in Phase One, or e) unique characteristics of the prototype tasks (which integrated reading, listening, and writing in a way that the TOEFL essays did not).

Decision Making Among Native-English-Composition Assessors

The seven highly experienced native-English-composition raters displayed fundamentally the same range of decision-making behaviors that the 10 ESL/EFL assessors in Toronto had when they rated TOEFL essays in Phase One. But the think-aloud protocols of each group differed in certain qualitative ways. One noticeable distinction was that most of the native-English-composition raters appeared to abstract their decision making into overall, reflective impressions about each essay. Rather than the step-by-step reporting and progressive decision making that the Toronto participants typically displayed as they read each essay, most of the U.S. raters tended to read a composition quickly, then step back and make evaluative reflections on the composition as a whole, summarizing the key ideas and the manner in which the writing was developed in conjunction with their impressions and judgments of it. For example:

> Number 62. Uh, it's the university topic ... Now, there are advantages and disadvantages. If the university comes to his community there are only advantages. Uh, the, the reasons, uh, "GOOD ENVIRONMENT FOR ..." this kind of thing. There, there's a vagueness, little pieces of somewhat unidiomatic, uh, possibly, uh, with problematic syntax. Very skimpy. One sees the, the, the general thought in the sense of what is implied, are relaxing, he would ha-, there's a relaxing environment, a healthful environment, I guess one would say. Good, there's good transportation, uh, there are good stores. The way the words, uh, express the ideas, uh, and the sparseness of them, the, there are only about 10 lines in the whole piece. There's too little undeveloped. It's very, it's incompletely expressed. I'd have to give this a 2 also. Number 62 is a 2. (Doug)

> The next essay is 128... It's on relatively short vacations, and the writer very well expresses the idea in the first paragraph that the short vacations are a better idea, then suggests that he or she will explain why. Suggests a vacation schedule, one which logically would provide "optimal school performance." And suggests re-, recreation in the proper doses will help students become better students and, and start new cycles of learning. Although they should not wait so long that they forget what they've learned in the previous cycle before they begin the new cycles ... The example of, uh, in the Ukraine, summer vacations that last from 12 to 14 weeks, uh, made it difficult for people to return to school ... The conclusion is a little brief, but generally the paper is well handled. I would say it's a low 5. Low 5. (Fred)

These brief, synoptic analyses of each essay typified the think-aloud protocols of most of the native-English-composition raters. But, as with the ESL/EFL assessors, there were variations from rater to rater in their assessment styles. Many of the native-English-composition assessors did display a certain amount of progressive decision making. Indeed, two of the U.S. raters, Lucy and Kathy, did so in a manner that closely resembled the assessment behaviors of the ESL/EFL assessors in Toronto, not only in their progressive decision making while they read (which tended to follow the prototypical sequences described above for Phase One), but also in their attention to aspects of English usage. For example:

> Paper number 28. "IN MY OPINION ..." I was thinking 4 but I'm wondering now. It's starting to disintegrate or deteriorate. Uh, "BECAUSE OF THIS ..." Lots of spelling errors. They may be typing errors. I'd suspect they're spelling errors. "THE FEMILLY ROOM ..." There we have a fairly important missing verb. "THE FAMILY MEMBERS ..." This is certainly not a 4. I'm going to say it's a 3 because, uh, I see the position. The position's fairly clear to me. Uh, but there are some, uh, problems in grammar and usage and mechanics here, and, uh, I would say they're, they accumulate, uh, and the accumulation of errors, things like missing verbs, important missing verbs, really begins to interfere with meaning. Uh, and at the same time I would say the development, uh, and reasoning of the family room kind of breaks down. Uh, that, tha-, certainly that third paragraph, there's an idea there but it's never brought up, out effectively and therefore I'm going to score number 28 a 3. (Lucy)

> I am now reading essay number 97. And looking at it, it seems relatively brief. It's three paragraphs. "IN MY OPINION ..." O.K., the student clearly focuses, tells us his or her point of view. Uh, we have an agreement problem with "it," but that's not a major problem. The next paragraph "History teach," subject-verb agreement. "THE STUDENTS ABOUT ..." Well, it looks like there is a punctuation problem there. "Literature also help," "also helps," subject-verb agreement problem. "THE STUDENT TO ..." "Carrier," and I am sure it's supposed to be "career." The student sets out in the opening sentence to do two things, to say that it's important to do both, and really only develops the first point, which is important to study history and literature. Coupled with the brevity and lack of development of that idea in paragraph two, and the persistent errors in agreement, I would have to give this paper no more than a 2. (Kathy)

As in the example above from Kathy's protocol, the other native-English-composition raters also focused on language issues in many of the compositions to which they gave low scores. Several of them cited, as Sam does below, "ESL problems," "ESL errors," or "ESL markers" as catch-all phrases for indicators that the papers they were rating had been written by people who were learning English. For instance:

> "WHEN I WAS ..." This is number 134. "SHORT VACATION MAKES ..." Again, you kind of know what they're saying but those, are some, you know, they're not really horrible ESL errors, but they are getting in the way of meaning. "I KNOW I ... [sighs] Got to go with a low 3, high 2 possibly, but more a low 3. (Sam)
>
> Paper 52 ... There's, uh, this is a challenge. I think finding meaning here means supplying meaning. I have to, really. There's subject-verb problems. It's unidiomatic, uh, things that are unclear. There seems to be jumbled, uh, language. Word choices are problematic. There seem to be omitted words. Uh, diction, syntax, idiom, too fractured, too unclear. Uh, one, one can find meaning here, but, but it's again, supplying it, and that, that's not a good thing. It could be 1-2. There's some discernible meaning behind the mess. But, uh, I guess because there are discernible ideas, if it's a 1-2, one could give it a 2. It's a very low 2. Number 52 is a low 2. (Doug)

Nonetheless, the native-English-composition assessors characteristically struck an even balance in the attention they paid to ideas, argumentation, and language in the essays. This balanced consideration of the writing samples seemed to distinguish them from the ESL/EFL assessors, who probably by virtue of their professional focus on language teaching, seemed inclined to direct their decision making more toward examinees' uses of language, per se. What the native-English-composition assessors seemed to do, in contrast — perhaps because of their professional orientations toward writing in a literary or professional sense — was to focus more distinctly on examinees as writers. That is, the native-English-composition raters appeared to appreciate how examinees wrote or developed their TOEFL essays, going out of their way to discover or imagine the approach a student had taken while composing. While the ESL/EFL assessors tended to envision the situation of the writer as that of a university student enrolled in courses, or as that of a language learner, the native-English-composition assessors tended to envision the writer's situation as one of engagement in the processes of composing a particular essay. For example:

> Paper number 36 ... O.K. there seems to be someone, uh, O.K., taking a cre-, sort of a creative writing response. He's an official of the city. It's an imaginary city and he has heard that the university is going to be built, and he's thrilled, and he talks about it, uh. So, that's an interesting kind of response. Uh, he goes on to give the positive reasons of, uh, you know, we have a friendly town, a diverse town, there's little crime, good weather. It sounds like a promotion that he's writing here, uh. Uh, he, he mentions the negatives: the transportation isn't good, and it would be a big change for the city. He's taking this from the point of view of, uh, I've received the notice of, uh, the plans to build the university here, and I think it's a great idea. These are the positives. There are some negatives, you know.

> There's going to be traffic and all that. But, then he wraps it up with "It would do more good than harm," uh, therefore, I'd recommend it. Uh, it seems well organized. There's an engaging voice. I believe there are good supporting points, given the direction, uh, of the paper. I would give this a 5. Paper number 36 would get a 5. (Doug)

> Next essay is 107 ... Uh, 107 is one of these real, uh, real problem situations where a writer begins and is unsure where to go, and then breaks down and sa- and says, uh, that, uh. This is an essay of 6 lines, and in the first 3 lines the student tries to set up the idea of science and mathematics as necessary. Though less important than literature and history, as I think. They want to say "the common sense of life." Uh, then it gets to the apology, "I AM SORRY..." Well, the, the problem is this person, you know, begins but then has so little confidence that the person cannot go on and complete even the effort at [sneezes], excuse me, the essay. Uh, 107 is a 1. And this is the sort of essay that I think we should pull out, and look at, and talk about with the test developers to see, uh, what sorts of prompts might work more advantageously with the test population that has such difficulty and, and maybe some sort of, some form of test phobia. (Fred)

As many of these examples demonstrate, the native-English-composition assessors used a rich array of vocabulary to describe their impressions of the essays, in contrast to the technical or professional terms of grammar, usage, and discourse organization that predominated with the ESL/EFL assessors. Some of their terminology was unique to the field of composition instruction (e.g., "comma splice," "undeveloped," "voice"), but much of it was also highly metaphorical and creative, including terms such as "artful," "sparkle," "crispness," "skimpy," "skeletal," "halting," "refreshing," "tepid," or "roughness." Doug even summarized his impressions of one essay by anthropomorphizing it: "It clears its throat and doesn't go any further." Because of the richness and interpretive quality of such language, we did not think it appropriate to undertake an analysis of the text characteristics that the native-English-composition assessors cited in their decision making that was analogous to one presented in Table 2 for the ESL/EFL assessors.

Interestingly, the native-English-composition raters did not seem, with the exception of Fred and Jean (who are discussed next), to devote much attention to establishing or revising their criteria for rating the compositions in the way that the ESL/EFL assessors did. Specifically, they did not engage in macrostrategies — such as sorting the essays into piles — as a means of clarifying their rating decisions. Rather, they all rated the compositions confidently, rapidly, decisively, and one by one, seemingly with an assuredness that arose from their extensive experience in doing this kind of composition assessment. Their deliberations over their procedures or criteria for rating were brief. For example, when Fred began to rate his second composition, he remarked, "I would say this paper is a strong 5. If I had read more papers, then I would have been able to see what I would do in that regard, and if I regard it as a stronger paper."

Apart from Sam, who described in detail in advance of reading the sample essays how he was going to use the SAT II: Writing test criteria to rate them, Jean was the only native-English-composition assessor in our sample who did report attending extensively to her criteria for judging the essays as well as to comparing the scores she gave to related compositions. (It may be worth noting that she was also the one native-English-composition assessor who rated herself as a "competent," rather than an "expert," assessor of writing.) For instance, near the end of her rating session, Jean puzzled over one composition, and felt she had to rationalize her criteria for scoring it in terms of her usual assessment practices while teaching:

> I'm on to 1-, uh, to number 137 now. It's a short three-paragraph, typed essay about the short vacations. I'm reading silently ... It takes a definite stance that short vacations are better. Uh, there's a paragraph on why they're better, there's a paragraph on why long vacations are not better, and then there's a conclusion. Uh, I could give it a 4. I do so rather reluctantly. Uh, I think for instance of the paper just preceding it, which I also gave a 4, which is clearly a little bit stronger, but you know, even within those ranges, there's going to be some variation in quality. There's going to be a high 4 and a low 4, and I think that's what I've got here. Uh, uh the previous one was a high 4, and this one's a low 4. But I still feel fairly secure about the 4. When I, when I grade papers, and I give students a grade, I understand that whatever score they get tells them, we hope objectively, where the paper has its strengths and where it has its weaknesses. So that for instance, uh, uh, I'll use the A, B, C categories here, just because I have them handy in mind and have been working on them with my own students. Uh, a C paper for instance will very rarely use any illustrations and it may have some problems with organization and approach, that it's logically not being completely clear. A B paper my students understand will be much stronger in logic than a C paper, but it still may not reach the illustration level, whereas an A paper will accomplish both things, perhaps rather neatly. (Jean)

In sum, the native-English-composition assessors seemed oriented toward practices, conceptualizations, and terms that displayed their professional, cultural backgrounds in English literature or composition departments at universities. They tended to rate the compositions more quickly, but also more creatively and reflectively, than the ESL/EFL assessors had, doing so in a manner that balanced considerations of ideas, rhetoric, and language use. Their transcribed protocols were briefer, by about half (both in terms of number of cassette tapes and total pages of transcribed data), than those of the ESL/EFL assessors, though they had each rated 40 essays, rather than the 60 essays the ESL/EFL assessors evaluated in Phase One. Their protocols showed most of them to be highly proficient in assessing writing and to be actively engaged in the assessment tasks. Like the highly proficient essay raters studied by Wolfe, Kao, and Ranney (1998), most of the present native-English-composition raters (except for Lucy and Kathy)

tended to distance themselves from the compositions when making their evaluative decisions, reviewing their impressions of the writing overall and in general terms, rather than processing their decisions progressively while they read.

However, some of these observations may be worth reconsidering in terms of the assessors' conditions of task performance. Unlike the sample of assessors in Toronto, who worked and were oriented to thinking aloud together, the native-English-composition assessors in our sample were sent materials by courier, performed the assessment tasks independently, without face-to-face demonstrations of thinking aloud, and were paid lump sums for their participation. It is therefore possible that some of their data, and the interpretations of it above, may be artifacts of these circumstances, rather than representations of fundamental differences between the native-English-composition and ESL/EFL assessors.

For example, some of the native-English-composition raters seemed to have produced think-aloud protocols that were more post hoc, retrospective reports of their thinking, rather than concurrent, progressive verbal reports (cf. Cohen, 1994; Smagorinsky, 1994). Except for Kathy and Lucy, they may have understood and enacted the think-aloud procedure in a different way than the participants in Toronto had. This may explain some of the reflective, retrospective qualities to their decision making, though it may also be that they were predisposed (or practiced, perhaps through extensive experience with holistic scoring) to assess compositions in this manner anyway. Similarly, performing the assessment tasks for a fixed stipend may have prompted the native-English-composition assessors to go through the assessment tasks more quickly, producing briefer and less verbose think-aloud protocols than the ESL/EFL assessors did, who were paid for their involvement at an hourly rate. Alternatively, the ESL/EFL assessors may have produced more extensive think-aloud reports, based on higher or different expectations, knowing they were performing for their professor as well as peers, and given their stakes in the research process itself.

Decision Making While Assessing Prototype Tasks

In assessing the TOEFL 2000 prototype writing tasks, the ESL/EFL assessors demonstrated the full range of decision-making behaviors they had previously displayed in Phase One of the research, as well as stylistic variations from rater to rater, as noted in previous analyses. But there were several notable, qualitative differences in their decision making while assessing these new tasks, compared to the TOEFL essays. First, the raters all seemed to expand their consideration of the writing prompts beyond that which they gave to the TOEFL prompts. This is probably because the prototype tasks were novel and more complex than the TOEFL essay prompts. While the TOEFL essays involved a simple, standard prompt to write an essay, the TOEFL 2000 prototype tasks differed from one another. Each had new, unique instructions

and expectations for particular genres of writing and involved complex combinations of content from reading or listening materials. These elements had to be comprehended carefully by a student completing the writing tasks as well as by the rater who was to evaluate whether a student had fulfilled the task requirements. For example, the assessors had to mull cautiously over each task to determine what it involved:

> The first one that I see is, uh, student B, task b. I'm going to check the prompt or the explanation of task b that we were given, uh, with our package. It says here "RESPONSE TO A ..." So the first page here is the conversation that they heard. I'm going to read this. I want to, want to know what they, what they heard on the tape. So, uh, two people are talking. "COME ON" the man says, "COME ON JULIE ..." O.K. That's the end, and there are four, uh, questions directly below the, uh, this conversation. So I'm going to look at the next page now. This is the one that obviously the student was given. Uh, they, at the top it says "PLEASE LISTEN TO A CONVERSATION ..." So they listened to it, and there are four multiple-choice questions. So I'm going to read those quickly because I want to know, I'll get an idea about how well students understood what they listened to ... (Marty)

As the raters continued reading the compositions, they had to assess carefully whether a student had fulfilled the task requirements:

> ... Again the verbs are in the present tense. So this person is continuing with, uh, what they started in their own first sentence, but not what, what they were given to start. It's not the first sentence that they were given to start the, uh, composition. But so far they seem to be expressing ideas okay. Uh, the next bullet, "JUDY SAYS THE ..." (Marty, later in the same protocol quoted immediately above.) So, this is supposed to be a summary. It has some evaluation statements towards the end here. And, also, the writer is taking a position in this argument. Obviously the writer sympathizes with the tenants. There are some basic errors here, uh, for example ... Uh, they're saying like ... First of all, the information is not completely correct, as I said. The writer said that the bill was sent to the home of, to the flat owner, and that is not the correct information. Uh, the owner doesn't want to pay for it, and she is threatening. So, that's another error here, sentence structure error. Spelling error, "of course." So, I think I would give this a 2 for the time being, uh, because the writer did not succeed in putting down some of the information. And, in fact, that's all he or she is asked to do, to summarize the conversation. So, I can't say that that part has completely failed. So I am putting down a 2. O.K., that was task d, by student M. (Gary)

> Now this one, student B, task b, and it's listening, summary of a listening, uh, activity, "Listening to a Conversation," and I'm reading the prompt again. O.K., now, with this task we have explicitly stated this is a "summary," while with the other tasks we have "response." So, it is easier, I think for the students to, for the test takers, to organize their writing according to some format, some required genre or format. I'm reading the writing of this person now ... O.K., I'm done reading now. Uhm, this summary is a very detailed summary. So the writer really had good comprehension, good listening comprehension and good memory. Uhm, but that's not what this task is about. The task is about writing a summary, and here we have a point-form-notes type of summary, which I don't think is what the task is asking for. I think that, uh, the task was writing, was asking for, uh, a a summary, not point-form notes. Now, uhm, I cannot speak of any development as topic development because this is not united by some idea, it's just pieces of writing. (Patricia)

Indeed, the raters had to assess the task itself, its relative level of difficulty, as well as whether the student's writing had accomplished its stated purpose and requirements. For instance:

> Now I'm going to rate student M on task d. Uh, summary of a conversation between two friends about a legal situation. And the prompt demands, the writing instructions demand a five-sentence writing piece. The student has written about 6 lines. It's very difficult to read the, the essay. Uh, but I see, and I, uh, can detect the student's ideas here ... I'm reading the essay ... Uh, I see a change of addressee in the student's writing. Sometimes the student uses "they" and sometimes the first-person plural pronoun ... Uh, it does make sense in a way, so it's, as I see it, it's relevant to the topic and, uh, it's coherent to some point, to some extent ... Uh, no organization, uh, no ... Uh, probably it's the prompt that does not require an organization. I don't know. Uh, I don't see many errors, either, uh, grammatical or lexical. But, uh, the student made good use of the conversation. (Ed)

> Uh, the response should be at least five sentences in length. And although as much as possible I'd like to give students the benefit of the doubt, this is just a big long list, and, uh, no attempt at writing complete sentences, so a bit of a cop-out in that way. So they've really only written two sentences. Uh, they, student, uh, ra- the student rather has basically got the gist of the conversation. I don't know that it's appropriate tone for a political science class. Uh, although I, I am not sure about [laughs] the the task. I am not sure how to interpret them, or whether tone in fact really matters, or whether just a straight summary is what is being asked for. (Jane)

In addition to assessing the appropriateness of the written text and its vocabulary, discourse appropriateness, and register, the raters were compelled to assess an examinee's understanding, or misunderstanding, of the prompt material. This behavior seldom appeared in their assessments of the TOEFL essays. For example:

> Let me try to figure out what she writes ... O.K., uh, I tried to see some of the words I can figure out, and it seems like she misunderstood the story. She got the point that the dishwasher was broken, then got it repaired. The owner was angry, but the, the, the relationship connec-, and, uh, O.K. the, uh, relationship between, among Paul and Sarah, or between the owner and Paul and Sarah became worse ... But it's not, she doesn't write about the legal, it doesn't look like she is clear with the legal points. And, I think this is, I don't know. Either 2 or 3. And the, the only way I can put the number would be, uh, through the, my judgment about their, her language use, because there are some obvious points where she didn't get the meaning from the reading, and then she is not doing 100%, uh, achieve-, she is not doing the, answering the writing instructions 100, for, for 100%. Uh, because her summary is not specifically on the legal situation. (Melissa)

> Uhm, the writer's target doing the task, the task is asking for an explanation of the problem at the beginning. So, the writer did do that or at least started doing that ... uhm, but the second part, making a request, is really very poorly done. And also the conventions of making a request are obviously not known to the writer. Uhm ... Here I see it as a sociolinguistic problem, how to express different functions in a new language, that is the problem I can see here with this writer. That's where the poor completion of the task is coming from. (Patricia)

A further, related concern was how examinees made use of the source materials provided in the prompts, and their appropriateness or effectiveness in doing so. For instance: there is no organization, but, uh, there are some ideas here as there were in the lecture. But the person is not accurate about "pitch," and ... And the, it seems the person tried to copy from what was in the lecture. So there's no creativity here ... And not, uh, not a good organization ... And I see most of the words are in the original lecture. But the student conveys some ideas and is right about "pause." (Ed)

> And the fifth sentence that the student has written, "SOME TIMES A ..." I, although my memory is not very good today, I am sure that wasn't mentioned in the lecture. Uh, let's see, sorry I just took, I'm sorry I'm just reading back through the lecture. Uh, yeah, the, uh, point that was made in the lecture was that with an amplifying device like a microphone, the speaker can use a natural tone, and that is something I just read from the lecture. Uh, so the fifth point, although the student clearly got the idea of microphone, uh, the, the point that the student has

made is off topic. So I'm looking back, and, uh, two and a half sentences are off topic, so it's not a very complete answer, I would say. (Jane)

Similarly, the raters often assessed an examinee's ability to make appropriate and creative use of the source materials, and whether the writer did so in a manner appropriate to the task demands:

> Uh, the student is trying to, uh, uh, put things in their own words, and I always appreciate that when they do it. Th- They're certainly not copying from the prompt, so that is good. Uh, but I think that, uh, there are serious problems, uh, both in the organization in that, uh, there's no framing in any way to the, uh, description. It's just a summary, point by point, of the, uh, conversation. (Scott)

> There is no trouble understanding any of these sentences. And, uh, it's expressed, I think, in a, an appropriate, at an appropriate level of politeness, which shows some skill. But they didn't close it with "thank you" or "yours sincerely," which shows a little, slight lack of knowledge about correct or, or suitable letter writing formats. (Marty)

A final observation is that — perhaps because of the way the tasks had been distributed to the assessors (i.e., as six tasks grouped by each ESL student) — some of the raters tried to compare students' abilities across the five prototype tasks, seemingly developing an image of what each examinee was capable of writing in English. This appeared to provide them with a fuller representation of each student's abilities than just a single essay sample did. However, it also involved considerations of multiple task performances, rather than the single rating done for the TOEFL essays. For example:

> O.K., uh, all this is a bit better than the other, uhm, writings of student Z. It's better organized ... I'm scanning the, the writing piece again, uh ... Well, this is, uh, this is better organized, coherent to a much larger degree than the previous ones, the previous writings of this particular student ... [sighs] uhm, however, it is not good. It's just better than the rest. (Patricia)

But some raters sometimes found it challenging to switch between the various tasks, and to keep each task's particular demands in mind. For instance:

> O.K., the student had to provide a summary. That's quite a different task than having to write a letter, uh, providing a problem or stating the problem and suggesting a solution. Anyway, it's quite interesting. (Jane)

In sum, the TOEFL 2000 prototype writing tasks appeared to provide the ESL/EFL assessors with a fuller, more complex image of each examinee's writing than they had obtained from just a single TOEFL essay. The various genres, registers, and sociolinguistic situations stipulated in the prototype tasks prompted the assessors to consider how examinees, in addition to demonstrating formal aspects of composing and argumentation, made use of source materials, fulfilled the requirements of specific writing tasks, and displayed their abilities effectively and appropriately.

Revisions to the Descriptive Framework

Based on the foregoing observations, we made various minor revisions to the wording and categories of the descriptive framework of decision-making behaviors in the process of coding the think-aloud data. As noted earlier, this led to a refinement of Table 1, which is presented as Table 3 following the list. This is the version of the descriptive framework that we used for coding (following the exemplars in Appendix E). Our revisions at this stage involved the following:

- moving the strategy "scan whole composition" to the self-monitoring/interpretation category and adding the strategy "observe layout" to the language focus/interpretation category to entail a wider range of rating behaviors here
- combining two self-monitoring/judgment strategies into one, "consider own personal response or biases"
- deleting the self-monitoring/judgment strategy "distinguish interactions between categories," because it overlapped with either the strategy "compare with other compositions or anchors" or "articulate or revise scoring decision"
- adding the self-monitoring judgment/strategy "articulate general impression," because raters did this with some frequency
- incorporating the rhetorical and ideational/judgment strategy "assess topic development" into "assess reasoning, logic, or topic development," because these behaviors tended to occur together and to be difficult to distinguish
- incorporating the rhetorical and ideational/judgment strategy "assess helpfulness in guiding readers" into either the strategy "assess reasoning, logic, or topic development" or the strategy "articulate general impression," depending on the focus of a rater's utterance

- adding a new rhetorical and ideational/judgment strategy "assess text organization" that raters displayed with some frequency

- deleting the language/judgment strategy "assess writing skill," because it was often impossible to distinguish from the language/judgment strategy "rate language overall" or the rhetorical and ideational/judgment strategy "assess style, register, or genre"

- deleting the phrase "command of" in several of the language/judgment strategies

After coding the think-aloud data from the prototype tasks, we concluded that three minor changes to the descriptive framework were necessary to account for raters' decision making while they assessed compositions written for these tasks (and which could be applied to the TOEFL essays as well). First, we expanded the first self-monitoring/interpretation strategy to "read or interpret prompt and/or task input," because, as noted in the previous section, the TOEFL 2000 tasks required raters to attend considerably more to the content and requirements of the writing prompts and task instructions, and in unique ways for each prototype task, than the TOEFL essays had. We also deleted the word "essay" from this strategy, because the writing tasks now included many different text genres. Second, we added reference to "discourse functions" in the rhetorical and ideational/judgment strategy "assess style, register, discourse functions, or genre," because certain new tasks, such as writing a note, required raters to judge the appropriateness and writers' uses of discourse functions, such as requesting or apologizing, in a way that the essay genre did not. Third, we added the new judgment strategy "consider use and understanding of source material," which the raters displayed frequently in assessing prototype tasks that required summarizing or references to reading or listening passages. At this point in the progress of the project, though, it was too late to incorporate these three revisions into the coding scheme, so they do not appear in Table 3 and do not feature into our statistical analyses.

Table 3

Revised Descriptive Framework of Raters' Decision-Making Behaviors While Scoring TOEFL Essays

Self-monitoring focus	Rhetorical and ideational focus	Language focus
	Interpretation strategies	
* read or interpret essay prompt * read or reread composition * envision personal situation of writer * scan whole composition	* interpret ambiguous or unclear phrases * discern rhetorical structure * summarize ideas or propositions	* observe layout * classify errors into types * edit phrases for interpretation
	Judgment strategies	
* decide on macrostrategy for reading and rating * consider own personal response or biases * define or revise own criteria * compare with other compositions or "anchors" * summarize, distinguish, or tally judgments collectively * articulate general impression * articulate or revise scoring decision	* assess reasoning, logic, or topic development * assess task completion * assess relevance * assess coherence * assess interest, originality, or creativity * identify redundancies * assess text organization * assess style, register, or genre * rate ideas or rhetoric	* assess quantity of total written production * assess comprehensibility * consider gravity of errors * consider error frequency * assess fluency * consider lexis * consider syntax or morphology * consider spelling or punctuation * rate language overall

Note. This version of the descriptive framework was used for coding data in Phase Two.

Because the decision-making behaviors of the native-English-composition assessors did not differ substantively from those of the ESL/EFL composition assessors, we did not make any revisions to the descriptive framework based on the former's decision-making behaviors. However, in reviewing the statistical trends reported in the next section of this report, we observed that certain decision-making behaviors occurred with so little frequency in our data that we think seven of them could reasonably be combined, for future purposes, with other categories of decision making that are logically similar to them. In particular, we have combined:

- "observe layout" with "scan whole composition"
- "decide on macrostrategy for reading and rating" with "compare with other compositions or anchors" and "summarize, distinguish, or tally judgments collectively"
- "interpret ambiguous or unclear phrases" with "edit phrases for interpretation"
- "assess task completion" with "assess relevance"
- "assess coherence" with "identify redundancies"
- "assess style, register, or genre" with "assess text organization"
- "consider gravity of errors" with "consider error frequency"
- "assess fluency" with "assess comprehensibility"

Table 4 displays the next, and final (for this project at least) revision of the descriptive framework. It incorporates the foregoing revisions to accommodate the prototype tasks, as well as statistical trends in the frequency of decision-making behaviors. Table 4 presents a descriptive framework of 27 decision-making behaviors that, on the basis of the research conducted in Phases One and Two of this project, we now believe accounts comprehensively for the decision-making behaviors that diversely experienced ESL/EFL assessors as well as experienced native-English-composition assessors reported during think-aloud protocols while assessing a broad range of TOEFL essays — essays written in response to different prompts as well as in response to a variety of prototype TOEFL 2000 writing tasks that incorporate reading and listening requirements.

Descriptive Statistics for the Coded Think-Aloud Data

We analyzed the coded think-aloud data to describe the frequencies (converted to percentages for each composition) of the decision-making behaviors found among the assessors, as well as to test for differences between the two groups of assessors (ESL/EFL raters vs. native-English-composition raters) and between the types of writing tasks (TOEFL essays vs. TOEFL 2000 prototype writing tasks). In this section, we report results for analyses of think-aloud protocols on the TOEFL essays first. Second, we examine differences between the ESL/EFL and native-English-composition assessors, and third, we look at differences between essays scored as high or low. Finally, we discuss decision-making trends observed in the rating of the TOEFL 2000 prototype tasks.

Table 4

Final Descriptive Framework of Rater's Decision-Making Behaviors While Scoring TOEFL Essays

Self-monitoring focus	Rhetorical and ideational focus	Language focus
	Interpretation strategies	
* read or interpret prompt and/or task input * read or reread composition * envision personal situation of writer	* discern rhetorical structure * summarize ideas or propositions * scan whole composition or observe layout	* classify errors into types * interpret or edit ambiguous or unclear phrases
	Judgment strategies	
* decide on macrostrategy for reading and rating; compare with other compositions; or summarize, distinguish, or tally judgments collectively * consider own personal response or biases * define or revise own criteria * articulate general impression * articulate or revise scoring decision	* assess reasoning, logic, or topic development * assess task completion or relevance * assess coherence and identify redundancies * assess interest, originality, or creativity * assess text organization, style, register, discourse functions, or genre * consider use and understanding of source material * rate ideas or rhetoric	* assess quantity of total written production * assess comprehensibility and fluency * consider frequency and gravity of errors * consider lexis * consider syntax or morphology * consider spelling or punctuation * rate language overall

Note. This version was revised and finalized *after* the coding of the TOEFL 2000 prototype tasks and the statistical analyses were completed.

Trends in decision making: TOEFL essays. Table 5 presents the means and standard deviations for all 35 decision-making behaviors (as outlined in Table 3 and exemplified in Appendix E) on the 20 TOEFL essays selected to represent all six possible score points for the two groups of assessors (106 think-aloud protocols from the 10 ESL/EFL raters, and 31 think-aloud protocols from the seven native-English-composition raters). It shows that the 35 behaviors we coded account comprehensively for the decision making that both the ESL/EFL and

native-English-composition assessors reported verbally over 137 think-aloud protocols while they rated the same 20 TOEFL essays. The 35 behaviors distributed fairly evenly across all raters, each behavior accounting for 10% or less of the raters' thinking about their assessments — except for the behavior "reading or rereading the composition," which accounted for about 20% of their decision making. We take this distribution to be evidence of the suitability of the revised descriptive framework (Table 3) to describe these raters' assessment behaviors, and thus, an initial empirical verification of its validity.

Several of the specific behaviors, however, proved to account for 1% or less of overall decision making, and so reasonably can be combined with other categories in the descriptive framework that are similar to them. Specifically, these behaviors (shown in Table 5) were:

- scan whole composition
- decide on macrostrategy for reading and rating
- interpret ambiguous or unclear phrases
- summarize, distinguish, or tally judgments collectively
- assess coherence
- assess style, register, or genre
- consider gravity of errors
- assess fluency

These revisions were presented earlier in Table 4, but we observe the unique importance of specific behaviors that appeared in Table 3, which we followed in our coding. For instance, deciding on a macrostrategy for rating does not occur very frequently, per se, but when it does, an assessor is making a major decision in the process of rating a set of compositions.

As Table 5 shows, the mean percentages of decision-making behaviors were mostly similar for the ESL/EFL assessors and for the native-English-composition assessors. Most coded behaviors account for about the same extent of decision making for both groups on the 20 TOEFL essays. We take this to be further verification of the relevance of the decision-making categories presented in Table 3. Notable differences, however, were that the native-English-composition assessors tended (proportionally) to devote slightly more attention to:

- discerning rhetorical structure
- summarizing ideas or propositions
- assessing reasoning, logic, or topic development
- assessing interest, originality, or creativity
- identifying redundancies

In contrast, the ESL/EFL assessors tended (proportionally) to devote slightly more attention to:

- classifying errors into types
- assessing quantity of total written production
- considering syntax or morphology
- rating language overall

These tendencies in the mean percentages shown in Table 5 corroborate our qualitative impressions of the think-aloud data, and probably reflect slightly different orientations in the practices and values of the respective professional fields from which the two groups of assessors were selected. The ESL/EFL assessors also (proportionally) devoted slightly more attention to reading or rereading the compositions and to articulating or revising their scoring decisions. This may reflect either their relative lack of experience, compared to the native-English-composition assessors, or as discussed earlier, it may have been that they approached the rating tasks with more deliberation and greater expectations than the native-English-composition assessors.

Differences between ESL/EFL and native-English-composition assessors. To more clearly view trends in these data, we aggregated them under the logical headings in the descriptive framework — that is, two types of strategy (interpretation and judgment) and three types of focus (self-monitoring, ideational and rhetorical, and language). Table 6 shows the same data as Table 5, but aggregates the 35 decision-making behaviors into a) grand means and standard deviations for all decision-making behaviors involving either interpretation strategies or judgment strategies, and b) grand means and standard deviations of decision-making behaviors that focus either on self-monitoring, ideas and rhetoric, or language. Viewed in this manner, Table 6 demonstrates that both the ESL/EFL and the native-English-composition assessors devoted approximately the same amount of attention either to interpreting (about 40% of their decision making) or to judging (about 60% of their decision making). The stability of these figures further verifies this aspect of the revised descriptive framework.

However, differences do appear in Table 6 between the two groups in terms of the attention they devoted to behaviors involving self-monitoring (ESL/EFL: $M = 44\%$; native-English-composition: $M = 38\%$), behaviors attending to rhetoric and ideas (ESL/EFL: $M = 19.6\%$; native-English-composition: $M = 33.6\%$), and behaviors attending to language (ESL/EFL: $M = 36.4\%$; native-English-composition: $M = 28.3\%$). Native-English-composition assessors appeared to devote more attention to rhetoric and ideas than the ESL/EFL assessors did. Looking at the native-English-composition assessors alone, the extent of attention they devoted to rhetoric and ideas appears similar to the attention they devoted to language. This supports our qualitative impression that native-English-composition assessors seemed to balance their attention to matters of rhetoric and ideas and matters of language fairly evenly while rating the TOEFL essays. In contrast, ESL/EFL assessors seemed to devote more attention to language overall than they did to rhetoric and ideas, and slightly more attention to self-monitoring behaviors, in their assessments of the TOEFL essays. As explained earlier, we were not able to assess whether any of these differences were statistically significant, because of the small sample size and the skewed distribution of essays from different points on the six-point scale.

Table 5

Mean Percentages and Standard Deviations for 35 Decision-Making Behaviors Among ESL/EFL and Native-English-Composition Assessors

Decision-making behavior	Combined raters *N* = 17 137 protocols *M* (*SD*)	ESL/EFL raters *n* = 10 106 protocols *M* (*SD*)	Native-English-composition raters *n* = 7 31 protocols *M* (*SD*)
Read or interpret essay prompt	1.9% (1.8)	2.4% (7.2)	0.3% (1.8)
Read or reread composition	19.9% (10.4)	20.4% (11.3)	17.9% (10.4)
Envision personal situation of the writer	1.3% (5.6)	1.1% (2.6)	2.1% (5.6)
Scan whole composition	0.4% (1.3)	0.4% (1.9)	0.2% (1.3)
Decide on macrostrategy for reading and rating	0.1% (0)	0.1% (.8)	0% (0)
Consider own personal response or biases	0.8% (3.3)	0.7% (1.8)	1.1% (3.3)
Define or revise own criteria	2.7% (3.7)	2.5% (4.4)	3.1% (3.7)
Compare with other compositions or "anchors"	1.2% (2.4)	1.4% (3.1)	0.7% (2.4)
Summarize, distinguish, or tally judgments	0.6% (3.1)	0.4% (1.7)	1.1% (3.1)
Articulate general impression	4.0% (4.5)	4.2% (5.9)	3.2% (4.5)
Articulate or revise scoring	9.9% (4.2%)	10.4% (7.6)	8.3% (4.2)
Interpret ambiguous or unclear phrases	0.5% (2.0)	0.5% (2.3)	0.4% (2.0)
Discern rhetorical structure	1.8% (4.6)	1.2% (2.6)	3.8% (4.6)
Summarize ideas or propositions	3.7% (7.4)	2.9% (4.4)	6.5% (7.4)
Assess reasoning, logic, or topic development	6.2% (7.4)	4.9% (4.8)	10.7% (7.4)

Table 5, continued

Decision-making behavior	Combined raters $N = 17$ 137 protocols M (SD)	ESL/EFL raters $n = 10$ 106 protocols M (SD)	Native-English-composition raters $n = 7$ 31 protocols M (SD)
Assess task completion	1.5% (3.6)	1.4% (3.4)	1.7% (3.6)
Assess relevance	1.1% (4.2)	1.0% (2.6)	1.4% (4.2)
Assess coherence	1.1% (2.5)	1.2% (3.4)	0.6% (2.5)
Assess interest, originality, or creativity	0.9% (3.9)	0.6% (1.7)	2.2% (3.9)
Identify redundancies	0.7% (6.7)	0.3% (1.2)	2.0% (6.7)
Assess text organization	2.6% (3.2)	2.6% (3.8)	2.3% (3.2)
Assess style, register, or genre	0.6% (2.1)	0.6% (1.9)	0.5% (2.1)
Rate ideas or rhetoric	2.2% (3.5)	2.3% (4.0)	1.6% (3.5)
Observe layout	4.9% (5.1)	5.1% (6.1)	4.6% (5.1)
Classify errors into types	5.0% (7.2)	5.6% (7.0)	2.8% (7.2)
Edit phrases for interpretation	1.7% (3.0)	1.8% (3.7)	1.3% (3.0)
Assess quantity of total written production	6.9% (6.4)	7.4% (7.5)	5.3% (6.4)
Assess comprehensibility	2.0% (3.4)	2.1% (4.8)	1.8% (3.4)
Consider gravity of errors	0.9% (3.4)	0.6% (1.9)	1.9% (3.4)
Consider error frequency	1.2% (3.1)	1.1% (2.4)	1.8% (3.1)
Assess fluency	0.4% (2.2)	0.4% (1.4)	0.4% (2.2)
Consider lexis	2.6% (5.4)	2.5% (4.3)	3.0% (5.4)
Consider syntax or morphology	4.4% (3.4)	5.0% (5.5)	2.0% (3.4)
Consider spelling or punctuation	1.4% (3.6)	1.6% (4.0)	0.9% (3.6)
Rate language overall	3.2% (3.9)	3.4% (4.4)	2.5% (3.9)

Note. These behaviors were demonstrated in 20 TOEFL essays selected to represent all six possible score points for the two groups of assessors (106 think-aloud protocols from the 10 ESL/EFL raters, and 31 think-aloud protocols from the seven native-English-composition raters).

Table 6

Grand Mean Percentages and Standard Deviations for 35 Decision-Making Behaviors, Aggregated into Types of Strategies and Types of Focus

Decision-making behavior	Combined raters $N = 17$ 137 protocols M (SD)	ESL/EFL raters $n = 10$ 106 protocols M (SD)	Native-English-composition raters $n = 7$ 31 protocols M (SD)
Decision-making behaviors aggregated into overall types of strategies			
10 interpretation strategies	41.1% (17.0)	41.4% (16.8)	39.9% (17.0)
25 judgment strategies	58.9% (17.0)	58.6% (16.8)	60.1% (17.0)
Decision-making behaviors aggregated into overall types of focus			
11 self-monitoring focus behaviors	42.6% (10.3)	44.0% (15.0)	38.0% (10.3)
12 rhetorical and ideational focus behaviors	22.7% (14.1)	19.6% (12.6)	33.6% (14.1)
12 language focus behaviors	34.6% (11.8)	36.4% (15.0)	28.3% (11.8)

Differences between essays scored high or low. One impression we had from Phase One was that the ESL/EFL assessors devoted greater attention to rhetoric and ideas when they rated TOEFL essays very high (i.e., 5 or 6), compared to essays they rated very low (i.e., 1 or 2). To the latter group, they seemed to devote relatively more attention to language matters. To analyze this impression, we considered a subsample of the think-aloud protocols they had generated. The subsample was composed of a) the 11 essays that had been scored consistently in Phase One as either 1 or 2, and b) the five essays that had been scored consistently in Phase One as either 5 or 6. We also considered the 20 protocols produced on these essays by the U.S. raters.

Table 7 presents the means and standard deviations in the subsample for all decision-making behaviors involving the three types of focus (self-monitoring, ideational and rhetorical, or language) for essays rated very low or very high by the ESL/EFL and the native-English-composition assessors. As a comparison of the grand means of these aggregated decision-making behaviors shows, ESL/EFL assessors tended to devote more attention to rhetoric and ideas on essays they rated very high ($M = 30.0\%$) compared to essays they rated very low ($M = 22.5\%$). But the extent of their attention to language remained consistent for the essays they rated very high ($M = 31.4\%$) compared to the essays they rated very low ($M = 31.9\%$), seemingly because they also did less self-monitoring of their own assessment behaviors on the essays they rated as high ($M = 38.6\%$), compared to the essays they rated as low ($M = 45.6\%$).

Native-English-composition raters behaved in a similar manner, as they devoted more attention to rhetoric and ideas in the essays they rated as high (M = 37.0%) compared to the essays they rated as low (M = 29.8%). But native-English-composition assessors devoted relatively more attention to language in the essays they rated as low (M = 32.3%) compared to the essays they rated as high (M = 26.0%). Unlike ESL/EFL assessors, though, native-English-composition assessors performed about the same amount of self-monitoring behaviors for the essays they rated as high (M = 37.0%) as for the essays they rated as low (M = 37.0%).

Table 7

Grand Mean Percentages and Standard Deviations for 35 Decision-Making Behaviors, Aggregated into Types of Focus, for TOEFL Essays Rated Very Low or Very High

Decision-making behavior	ESL/EFL raters n = 10 83 protocols M (SD)	Native-English-composition raters n = 7 20 protocols M (SD)
11 Self-monitoring focus behaviors		
Essays rated low (1 or 2)	45.6% (14.0)	37.8% (8.9)
Essays rated high (5 or 6)	38.6% (12.1)	37.0% (5.5)
12 Rhetorical and ideational focus behaviors		
Essays rated low (1 or 2)	22.5% (14.3)	29.8% (14.1)
Essays rated high (5 or 6)	30.0% (8.0)	37.0% (10.9)
12 Language focus behaviors		
Essays rated low (1 or 2)	31.9% (14.3)	32.3% (10.9)
Essays rated high (5 or 6)	31.4% (8.9)	26.0% (9.3)

Trends in decision making: prototype tasks. Table 8 presents means and standard deviations for the 35 decision-making behaviors that the seven ESL/EFL instructors displayed in the 210 think-aloud protocols they generated while rating the five TOEFL 2000 prototype writing tasks. The resulting data are similar in many respects to those that appeared in Table 5 for the TOEFL essays. This overall similarity supports both the viability of the new tasks (in the sense that they elicit patterns of assessment behaviors that resemble those demonstrated for the TOEFL essays) and the descriptive framework (in its being able to document these behaviors).

Some noticeable areas of difference appeared, however. The most obvious of these is that the ESL/EFL assessors tended to increase their decision-making behaviors related to "assessing

task completion" on all five TOEFL 2000 prototype tasks (from *M* = 1.4% on the TOEFL essays to *M* = 5.4% to 7.2% on the prototype tasks), either because they perceived this to be more important or salient in the new tasks, or because they were uncertain about how to score them. In support of the former interpretation, the ESL/EFL assessors also increased their attention to "assessing relevance" (to *M* = 5.2% on Task A and *M* = 2.6% on Task C, compared to *M* = 1.0% on the TOEFL essays) and "assessing interest, originality, or creativity" (to *M* = 2.9% on Task A and *M* = 4.1% on Task C, compared to *M* = .06% on the TOEFL essays) on prototype compositions requiring summaries. For prototype tasks requiring brief notes, the ESL/EFL assessors also increased their attention to "assessing style, register, or genre" (to *M* = 4.2% on Task B and *M* = 5.5% on Task D, compared to *M* = 0.6% on the TOEFL essays). Correspondingly, the ESL/EFL assessors lowered the extent of their attention to "assessing the quantity of total written production" on all of the prototype tasks (to *M* = 1.7% to 4.4%, compared to *M* = 7.4% on the TOEFL essays), perhaps because they found this of lesser importance in these types of tasks. Similarly, they reduced their attention to "spelling or punctuation" slightly on the prototype tasks (*M* = 0 to 0.4%, compared to *M* = 1.6% on the TOEFL essays), either finding this less of a concern than in the TOEFL essays, or perhaps because examinees demonstrated few problems of this order on the prototype tasks. On the latter point, it is worth noting that English proficiency seemed somewhat higher on the responses to the prototype tasks than the average English proficiency on the sample of TOEFL essays, which may have affected these results.

Table 9, which considers the decision-making behaviors displayed for the prototype tasks in aggregate form, puts these trends into a broader perspective. The ESL/EFL assessors showed about the same profiles for the distribution of their interpretation strategies (*M* = 40.4% to 45.0%) and their judgment strategies (*M* = 55% to 59.2%) on the prototype tasks as they had on the TOEFL essays (*M* = 41.4%, *M* = 58.6%, respectively). Considering the focus of their decision making, the ESL/EFL assessors remained about the same in their extent of self-monitoring on the prototype tasks (*M* = 43.0% to 44.9%) compared to the TOEFL essays (*M* = 44.0%). But the proportion of their attention to rhetoric and ideas in the prototype tasks increased distinctly (*M* = 25.9% to 28.6%), compared to that which they had exhibited for the TOEFL essays (*M* = 19.6%). Correspondingly, the proportion of their attention to language in the prototype tasks decreased (*M* = 26.9% to 30.5%), compared to the TOEFL essays (*M* = 36.4%).

Three explanations for this latter difference seem possible. First, the nature of the TOEFL 2000 prototype tasks may have prompted the ESL/EFL assessors to increase their attention to rhetoric and ideas, compared to language, relative to the TOEFL essays, leading them to behave in their decision making more like the native-English-composition raters in this regard. Perhaps they perceived there to be correct information or text types required for the prototype tasks, so they devoted less attention to deciphering examinees' meanings than in the more open-ended TOEFL essays. Second, since the proficiency of the ESL students who had produced the responses to the prototype tasks was higher on average than was demonstrated in the TOEFL

essays, the ESL/EFL assessors may have attended more to their rhetoric and ideas (following the trends displayed in Table 7 and our discussion of them). Or third, the ESL/EFL assessors, all but one of whom had participated in both phases of the study, may have increased their abilities for (or awareness about) assessing compositions through the processes of the research, possibly learning to attend more to rhetoric and ideas than to language in the prototype tasks, which they rated in Phase Two, compared to the TOEFL essays they had rated in Phase One.

Table 8

Mean Percentages and Standard Deviations for 35 Decision-Making Behaviors of ESL/EFL Assessors on Five TOEFL 2000 Prototype Writing Tasks

Decision-making behavior	Task A *M* (*SD*)	Task B *M* (*SD*)	Task C *M* (*SD*)	Task D *M* (*SD*)	Task E *M* (*SD*)
Read or interpret writing prompt	8.5% (6.3)	9.5% (7.3)	8.5% (6.5)	8.0% (6.6)	7.2% (5.6)
Read or reread composition	16.6% (9.0)	14.5% (8.0)	16.1% (9.9)	17.2% (9.1)	18.6% (9.8)
Envision personal situation of writer	1.2% (2.1)	1.4% (2.9)	2.2% (3.3)	1.5% (3.0)	2.6% (3.0)
Scan whole composition	0.5% (1.6)	0.4% (1.4)	0.6% (1.5)	0.3% (1.2)	0.8% (1.8)
Decide on macrostrategy for reading & rating	1.1% (3.6)	0.2% (1.0)	0.1% (0.4)	0.2% (0.9)	0.3% (1.0)
Consider own personal response or biases	1.5% (2.1)	1.1% (2.4)	1.1% (2.2)	1.4% (2.2)	1.5% (2.3)
Define or revise own criteria	3.9% (4.0)	4.3% (3.8)	3.6% (4.7)	3.3% (3.1)	3.9% (3.3)
Compare with other compositions or “anchors”	0.8% (1.9)	1.2% (2.4)	0.8% (2.1)	1.7% (2.9)	1.0% (2.0)
Summarize, distinguish, or tally judgments	0.3% (1.3)	0.3% (1.0)	0 (0)	0.4% (1.2)	0.3% (1.1)
Articulate general impression	2.6% (3.7)	2.1% (3.2)	2.3% (3.6)	2.2% (4.0)	1.8% (2.3)
Articulate or revise scoring	7.7% (4.6)	8.0% (3.7)	8.6% (4.7)	7.4% (3.3)	6.5% (4.6)
Interpret ambiguous or unclear phrases	0.4% (1.2)	0.9% (2.6)	0.3% (1.3)	0.5% (1.6)	0.8% (1.6)
Discern rhetorical structure	1.8% (4.2)	1.9% (3.6)	0.9% (2.2)	1.2% (2.8)	2.5% (3.4)
Summarize ideas or propositions	1.2% (2.2)	1.4% (3.4)	1.8% (3.1)	0.8% (1.9)	3.2% (3.3)
Assess reasoning, logic, or topic development	3.4% (3.4)	4.7% (4.4)	5.6% (5.0)	3.9% (3.9)	4.8% (4.7)
Assess task completion	5.4% (5.0)	7.2% (5.9)	6.0% (6.3)	5.8% (5.7)	5.6% (5.8)
Assess relevance	5.2% (6.4)	0.9% (2.3)	2.6% (3.9)	1.7% (2.7)	3.6% (5.3)
Assess coherence	0.5% (1.8)	1.0% (2.7)	1.0% (2.7)	1.7% (2.9)	0.6% (1.7)
Assess interest, originality, or creativity	2.9% (4.9)	2.3% (3.7)	4.1% (5.5)	2.3% (4.1)	1.5% (2.7)
Identify redundancies	0 (0)	0.3% (1.2)	0.7% (2.6)	0 (0)	0.7% (2.0)
Assess text organization	2.8% (3.7)	2.7% (3.3)	2.5% (3.6)	1.8% (3.2)	3.0% (3.5)
Assess style, register, or genre	1.3% (2.5)	4.2% (4.2)	1.0% (2.4)	5.5% (5.7)	0.6% (1.6)
Rate ideas or rhetoric	1.1% (2.1)	1.0% (2.5)	1.1% (2.3)	0.7% (2.1)	1.8% (2.6)
Observe layout	4.1% (6.3)	4.3% (5.0)	4.2% (6.3)	3.2% (5.0)	4.8% (5.6)
Classify errors into types	5.7% (4.7)	5.8% (5.3)	5.0% (5.1)	6.2% (4.8)	3.6% (3.9)
Edit phrases for interpretation	2.7% (4.1)	1.5% (2.5)	1.3% (3.0)	1.5% (2.5)	0.9% (1.9)
Assess quantity of total written production	1.7% (2.3)	1.8% (2.9)	4.4% (6.0)	2.4% (3.1)	2.8% (2.8)
Assess comprehensibility	2.7% (4.4)	1.9% (3.0)	2.0% (3.0)	2.9% (3.8)	2.6% (3.0)
Consider gravity of errors	1.3% (2.5)	1.2% (2.4)	0.4% (1.3)	1.4% (2.8)	0.9% (1.7)
Consider error frequency	0.7% (1.9)	1.4% (2.9)	1.0% (2.1)	1.0% (2.1)	1.8% (2.6)
Assess fluency	0.5% (1.4)	0.8% (2.3)	1.1% (2.7)	0.8% (2.2)	0.8% (2.5)
Consider lexis	1.5% (2.6)	2.1% (3.8)	1.7% (2.9)	2.0% (3.6)	2.5% (3.1)
Consider syntax or morphology	3.6% (4.1)	3.3% (3.8)	4.0% (4.3)	4.1% (4.5)	2.7% (3.5)
Consider spelling or punctuation	0.4% (1.2)	0.3% (1.3)	0 (0.2)	0.1% (0.5)	0.3% (1.0)
Rate language overall	4.3% (3.7)	4.3% (3.8)	3.4% (4.4)	4.8% (3.7)	3.1% (3.6)

Note. Task A = Listen to a lecture (on voice or geology); write a summary. Task B = Listen to a conversation (about TAs or election); write a note or a summary. Task C = Read a passage (on beads or urbanization); write a summary. Task D = Read a conversation (about a landlord); write a note or a summary. Task E = Read a lecture (on urbanization or diet); write a response.

Table 9

Grand Mean Percentages and Standard Deviations for 35 Aggregated Decision-Making Behaviors of ESL/EFL Assessors on Five Prototype Writing Tasks

	Task A *M* (*SD*)	Task B *M* (*SD*)	Task C *M* (*SD*)	Task D *M* (*SD*)	Task E *M* (*SD*)
Decision-making behaviors aggregated into overall types of strategies					
9 interpretation strategies	42.8% (13.1)	41.7% (13.6)	40.8% (15.9)	40.4% (14.6)	45.0% 15.0)
26 judgment strategies	57.2% (13.1)	58.3% (13.6)	59.2% (15.9)	59.6% (14.6)	55.0% (15.0)
Decision-making behaviors aggregated into overall types of focus					
11 self-monitoring focus behaviors	44.9% (14.2)	43.0% (11.4)	43.9% (12.3)	43.6% (13.5)	44.5% (12.9)
12 rhetorical and ideational focus behaviors	26.0% (13.3)	28.4% (11.3)	27.6% (14.1)	25.9% (9.2)	28.6% (13.1)
12 language focus behaviors	29.1% (13.4)	28.6% (9.2)	28.5% (10.0)	30.5% (10.6)	26.9% (12.6)

Note. Task A = Listen to a lecture (on voice or geology); write a summary. Task B = Listen to a conversation (about TAs or election); write a note or a summary. Task C = Read a passage (on beads or urbanization); write a summary. Task D = Read a conversation (about a landlord); write a note or a summary. Task E = Read a lecture (on urbanization or diet); write a response.

Assessors' Impressions of Prototype Tasks

Three recurring impressions of the TOEFL 2000 prototype tasks appeared in the interviews the research team in Toronto conducted among itself and in our inspection of the think-aloud protocols. These centered on a) the value of a range of writing tasks that are authentically integrated with reading and listening materials, b) the need for explicit instructions, criteria, and accounts of the conditions for performing each task, and c) the handling of the differences and comparability of the various tasks. These impressions highlighted features of the prototype tasks already described above, and indicated ways in which future prototype tasks might be improved. Our impressions have both positive and negative aspects, recognizing that the prototype tasks advance writing assessment in new directions but at the same time pose new challenges and uncertainties.

First, participants were highly positive about the move toward "authenticity" in the prototype tasks, particularly the integration of writing with listening and reading components, as well as the inclusion of a range of different text genres (other than argumentative essays, as in

TOEFL essays). Remarks such as the following appeared in various places throughout the think-aloud protocols and post-assessment interviews:

> That seems like a very personal question, which is nice. (Marty)
>
> The contents of the listening input is justifiable; students need to know this stuff. (Gary)
>
> With the letter writing task, the students' minds are more focused. (Scott)
>
> A problem in the past has been giving students ample content; these tasks provide them with enough information to help them express themselves. (Ed)

In addition to offering examinees opportunities for personal expression, the assessors found the summary tasks a) to be particularly germane to university-bound ESL/EFL students, b) to be more realistic than essay writing in their uses of English for academic purposes, and c) to be in accordance with recent research and theory demonstrating the interrelations of writing with reading and aural skills in academic settings. Stipulating the input, ideational content, and genre requirements for the prototype writing tasks helped to provide the ESL/EFL assessors with firm criteria against which they could judge whether an examinee had, or had not, successfully produced a text that fulfilled the stipulated task. For example:

> But the main problem is that the, what, what has been put down here, uh, is not anything that seems to be in the lecture. I am looking at the prompt now, and it seems to me that the lecture itself is pretty technical in the sense that it talks about this particular type of waves, seismic, seismic waves. The P-waves and the S-waves. There is no mention of that in the writing. Hmm, there is also a discussion about the detection of waves and, uh, and the structure of the earth. Uh, here, actually, it says nothing about earthquakes in this passage. (Gary)

But at the same time, while assessing the prototype tasks the raters often found themselves uncertain of exactly what the criteria for assessing a particular written composition should be, not only in terms of fulfilling expected goals for the task, but more particularly in respect to the register of language or vocabulary, as well as the formats of discourse, that might be appropriate to the tasks. For example:

> I'm dithering because it says for a law school class. I'm wondering if I should be looking for a more formal, uh, tone than what is written. (Jane)
>
> With this task we have explicitly stated, this is a summary, while with the other tasks we have "response." So, a response, to me, if I were writing this, I wouldn't know what format would be expected. (Patricia)

This is uh, not an essay task in a way. Uh, I'm not sure to what extent I should judge the student for sociocultural inappropriacies. (Scott)

This dilemma highlights the most frequent point that the ESL/EFL assessors raised during their interviews about the prototype tasks. They felt strongly that the instructions, criteria for assessment, and conditions for writing each task need to be specified in explicit, unambiguous detail and in a clearly articulated, conventional form — both for the examinees to set their goals for writing and for raters to be able to make informed judgments of the writing. This important point reiterates a suggestion made already in Cumming, Kantor, Powers, Santos, and Taylor (2000): A range of particular factors must be specified in designing and presenting prototype writing tasks for assessment. The ESL/EFL assessors felt they needed:

1. Precise knowledge of what the conditions for writing each task were. For example:

   I'm wondering if there was any kind of time limit for those ones. I don't know. (Marty)

   I am not sure whether or not, uh, the student was allowed to take notes. (Jane)

2. Guidelines on how to judge examinees' uses and comprehension of source materials from reading or listening prompts. For example:

   It looks, uh, suspiciously similar to what was in the reading. (Jane)

   I wasn't sure how much to penalize them for their unsuccessful comprehension of the reading or listening input. The reading texts influence students' writing. They borrow words and sentences from the prompt texts, and borrowing is an important strategy. But raters will need some guidelines to decide the difference between strategic borrowing and plagiarism. (Melissa)

   This is a comprehension question; if some students don't get the text, it will affect their writing, and this is not fair. (Ed)

3. Guidelines and exemplars to specify the qualities of writing expected for each task. For example:

   There should be more explanation about how the task will be rated and they should provide examples of what is being asked for, for example, of a brief note. (Ed)

   I didn't know what a summary and a note should really look like. (Melissa)

> Often these brief summaries can be produced without using very sophisticated language. (Scott)
>
> The letter is written, but not an ideal level, not at the ideal level. It need, needs more sophisticated writing, sophisticated writing, meaning with a little more support and reasoning. But this prompt does not generate that. (Melissa)

Second, criticisms were raised about certain instructions for the TOEFL 2000 writing tasks. As Patricia put it in her interview, "some of the instructions were ambiguous and vague," reiterating the need for more explicit guidelines, as indicated above. But certain comments also implied the need for clarity and purpose in the preparation of the prompt materials, and careful editing of them:

> The question they had to answer departs considerably from the overall tone of the reading passage. The reading passages were not written to address the issues that students are asked to write about. That makes them difficult. (Patricia)
>
> I don't remember any good essays on the "urban-rural pattern" question. They didn't have to read the passage to answer the question. Based on their answers, they mostly didn't refer to the passage. (Jane)
>
> I find that a little confusing. I, I understand now looking, looking ahead, I understand what the students have to do, but I find that task a little confusing because the first paragraph in their directions is written in the third person, and the second one is written as a directive, telling you, telling the writer to take the place of the, of Cathy and Stan. I found that confusing. (Marty)

Most of the assessors found the specification in one task — for examinees to "write five sentences" — problematic:

> The "write five sentence" direction was confusing and misleading, causing some students to write in point form. (Scott)
>
> They're repeating the same idea again and again, maybe to bring it up to the five sentences. (Jane)
>
> It might be better to say a minimum number of words, rather than five sentences. (Gary)

The third key point about the prototype tasks concerns the raters' handling of the unique demands of each task as well as their consideration of the set collectively. Interestingly, the assessors observed differences in difficulty among the various writing tasks. Moreover, there was a natural inclination to compare the tasks with each other, as well as to consider collectively

individual examinees' performance on all six tasks. But without explicit guidelines and information on the design or purposes of the tasks, the assessors were uncertain what to make of this while interpreting or scoring each task:

> There was a great discrepancy between the different tasks in terms of level of difficulty. The conversation texts were so much easier than the lecture texts. (Marty)

> This is the one about summarizing the conversation, so it's not the same. I thought the letter was a much more difficult task than a simple summary of the conversation. (Gary)

> The geology lecture was more difficult and less organized than the voice lecture. (Scott)

They also expressed concerns about potential biases in the input material, believing that the inclusion of certain topics may favor or disadvantage examinees who know the topic well, have certain sociocultural knowledge, or have certain styles of reasoning. For example:

> The geology lecture favored those who knew geology. (Gary)

> Although it's typical in TOEFL, there were biases in the dialogues, for example, in referring to school elections and teaching assistants. (Melissa)

> For students from more moralistic cultures, the solution to the problem might be dealing with Jack instead of getting a new room. (Gary)

Some of the assessors said they did not know how much knowledge they should expect examinees to have about social and institutional aspects of university life in North America. The interestingness and personal relevance of particular tasks or input materials also was an issue. For instance:

> The diet text was interesting and was easier than the other texts. (Marty)

> Some of the input passages were so boring and common that there was nothing interesting to say about them. They didn't draw out any personal experiences or background knowledge that much. (Patricia)

Procedurally, the assessors were also sometimes perplexed about how to handle the range of different types of writing, each with somewhat different expectations, lengths, and difficulty. This challenged their abilities to keep themselves focused while rating. For example:

> I'm finding it a bit difficult, uh, rating all these different tasks. It's so much easier if you're doing a whole lot of one kind of task. What I'm finding I'm having to switch gears and, uh switch gears and thinking of what to rate these. (Jane)
>
> The organization of tasks could be in terms of output or genre, instead of according to combinations of skills. (Gary)

Such potential confusions point toward a need for explicit scoring procedures for the TOEFL 2000 examination. These should indicate to raters exactly how they should score and handle not only individual writing tasks, but also an entire set of compositions and the differing demands and expectations of each task, compared to one another and in terms of assessing an individual's overall writing abilities.

5. Recommendations

Collectively, the findings from both Phases One and Two of this project have the potential to inform three major aspects of the design of the writing component of the TOEFL 2000 exam. Accordingly, we recommend that the next phases of research and development in this process address the three priorities that follow.

1. *Develop scoring rubrics and procedures for TOEFL 2000 writing tasks based on the descriptive framework of decision-making behaviors (Table 4).*

The descriptive framework, which was refined through successive phases of the present research, has sufficient grounding now to begin to serve as a basis for specifying scoring criteria and procedures for TOEFL 2000 writing tasks. The findings presented in this report not only document the complex decisions that experienced assessors tended to make and think about when they interpreted and judged TOEFL essays and a range of prototype tasks, they also describe how they did this. The next step is to use this information to prepare appropriate methods for scoring TOEFL 2000 writing tasks in a consistent, purposeful, and meaningful way.

The findings of this study point toward the value of a uniform, general rubric and scoring procedure that would encompass all TOEFL 2000 writing tasks, as well as additional specific, primary-trait scoring criteria unique to each task type or genre of writing presented in the test. That is, there are at least 27 specific decision-making behaviors that we can expect experienced composition assessors to make while they evaluate any writing sample of the kind likely to appear in the TOEFL 2000 examination. These should be accounted for explicitly in an analytic scheme that directs raters how to score TOEFL 2000 writing tasks generally and holistically. Preparation of a general procedure or scoring rubric could simply involve conversion of the categories and terms that appear in Table 4 into a set of procedural steps or benchmarks capable of guiding raters through their assessments.

However, participants in Phase Two of the present study also found they needed explicit guidelines to know how to evaluate examinees' performance on specific aspects of integrated tasks or unique text genres. This suggests the need for scoring guidelines particular to each of these that would account, for example, for such considerations as examinees' comprehension and uses of source materials, discourse or lexical features expected to appear in compositions, or the required format of a written text. Preparation of specific scoring rubrics particular to each task type or text genre (i.e., in the manner of primary-trait scoring methods) could be based empirically on analyses such as those undertaken to prepare the list of text characteristics to which raters attended (Table 2). There appear to be good reasons to make such scoring rubrics known and available in advance, both to examinees and to raters, because assessors participating in the present study frequently observed that they had to know what examinees had been instructed to do, and therefore should be perceived to be aiming to do, in their writing in order to be able to evaluate it effectively.

Criteria for expected levels of writing and scoring performance also need to be devised and validated. This is probably the major challenge for research following from the present study. The present findings suggest that such criteria for scoring should balance equal attention to interpreting and to judging key aspects of written compositions, as well as equivalent attention to rhetoric and ideas and to language features, as the most experienced of our participating assessors did. But there appear to be reasons to weight criteria more heavily toward language matters at the lower end of scoring criteria, and more heavily toward rhetoric and ideas at the higher end, because of the ipsative nature of the data we collected (which displayed these tendencies). That is, examinees seem to have to attain certain threshold levels of language proficiency before assessors attend thoroughly and sincerely to their ideas and rhetorical skills.

The descriptive framework can serve, in turn, as a basis for evaluating in pilot projects how raters make use of new scoring rubrics and procedures — which may be developed, for example, as checklists of desirable behaviors for raters to utilize or undesirable behaviors to avoid — and to ensure that a new scoring rubric elicits these behaviors effectively, fairly, and efficiently. A final suggestion is to develop scoring procedures for TOEFL 2000 writing tasks that prescribe how assessors will move through the sequence of making their assessments. For example, procedures should state which sequences assessors should follow for differing writing tasks.

Limitations to the descriptive framework also need to be borne in mind. The chief limitation is the extent to which the specified behaviors overlap with one another, either a) logically or categorically (e.g., are raters' behaviors interpreting unclear phrases related to *ideas and rhetoric* or to aspects of *language*, such as vocabulary, syntax, spelling, or handwriting, or to both?) or b) procedurally (e.g., these behaviors occur concurrently and very rapidly during rating, so are difficult to distinguish analytically). Although Phase Two of the present research addressed these problems and attempted to resolve some of their ambiguities, they are likely to remain a logical and practical limitation in future uses of the descriptive framework, simply because of the complexity of human assessment, the interrelations of language, ideas, and rhetorical forms, and the difficulty of distinguishing, if only for the purpose of analysis, discrete human behaviors from their holistically integrated nature. In addition, this research has shown that the terminology used to describe writing qualities, scoring criteria, and rating behaviors is inherently complex, open to variable interpretations, and difficult to specify. Although the present research team reached a consensus on the terms appearing in the descriptive framework, these terms are probably prone to differing interpretations among other populations as well as to differing realizations in particular cultural contexts and with specific text types.

2. *Acknowledge background influences on assessors' behaviors and scoring criteria, and develop a system for selecting, orienting, monitoring, and training raters in respect to these.*

Scoring written essays is a fundamentally interpretive and judgmental activity, based on prevailing norms of educational practice as well as individuals' past experiences. The very people who are well qualified to assess compositions also bring with them the background knowledge and skills that enabled them to gain such qualifications. Moreover, as the present research has demonstrated (cf., Kobayashi & Rinnert, 1996; Mendelsohn & Cumming, 1987; Song & Caruso, 1996), standards for rating ESL/EFL writing vary somewhat among differing cultural groups — even those as closely related as assessors experienced with English in second-language, foreign-language, and native-language settings — despite many commonalities in the criteria and rating behaviors they utilize. The profile questionnaire we developed for this research established that experienced composition assessors readily recognize certain influences on their scoring behaviors, such as those deriving from their past experiences with specific evaluation schemes, while teaching, or with their own writing or learning. This questionnaire instrument (or parts of it) could be used as a basis for specifying criteria to be used for selecting suitable raters for future research projects and for scoring TOEFL 2000 writing tasks. Some questionnaire items may also be helpful in monitoring influences that, on an ongoing basis, may affect raters' scoring. For example, some of the ESL/EFL assessors who participated in both phases of the present project were able to identify the influence of Phase One research activities on their performance in Phase Two assessment activities.

More specifically, findings from Phase Two of this study point toward specific decision-making behaviors that might predictably vary among composition assessors, either within or across certain professional fields (such as ESL, EFL or native-English composition) or among types of assessment tasks. Relatively high standard deviations from group means — for example, as shown in Table 5 — may indicate that decision-making behaviors that are potentially variable include the extent to which raters:

- read or interpret a prompt and/or task input
- envision the personal situation of a writer
- decide on macrostrategies for reading and rating
- compare compositions with other compositions or anchor papers, or summarize, distinguish, or tally their judgments collectively
- interpret or edit ambiguous or unclear phrases for interpretation
- discern rhetorical structure

- summarize ideas or propositions
- assess relevance and task completion
- assess coherence and redundancies
- assess interest, originality, or creativity
- assess text organization, style, discourse functions, or genre
- consider uses and understanding of source materials
- classify errors into types
- assess comprehensibility and fluency
- consider the gravity and frequency of errors
- consider lexis
- consider spelling or punctuation

Some of these behaviors may be more or less desirable, and worth encouraging or discouraging, for specific assessment purposes. As much previous research has shown, composition assessors can practice and be trained to conform consistently to a particular scoring rubric, and the extent of that consistency may be improved with the specificity and perceived value of a particular scoring rubric and assessment situation. A further point is that the diversity of raters who participated in this research raises the question of determining which types of educational qualifications and experiences should be valued in selecting raters of TOEFL writing tasks — particularly whether cultural diversity, certain academic disciplines, or specific orientations to writing should be prized and sought in the process.

One might, as Pula and Huot (1993) have suggested, try to develop and evaluate a model of the influences of personal backgrounds, professional training, cultural values, and work experience on assessor's rating performance (cf. Erdosy's [2000] subanalyses of the present research data). But doing so in detail probably only makes sense within the confined discourse community of a particular program at a specific educational institution, where such ongoing influences can be identified and monitored readily, rather than in reference to the large, diverse community of North American university contexts for which the TOEFL assessment is relevant, or the diverse community of raters who assess its writing components. All the same, such influences are worth attending to systematically in future TOEFL 2000 research and testing practices, as well as in the selection and training of composition raters.

3. *Continue to develop and refine prototype writing tasks along the lines already taken, but attend carefully to specifying the instructions, conditions, and expectations for each task type as well as their respective relations to one another.*

A number of findings from Phase Two of this study endorse the directions being taken in developing prototype writing tasks for the TOEFL 2000 examination, as well as in evaluating their feasibility from the viewpoint of assessment. Chief among the positive attributes of these prototype tasks are the variety of written genres they sample from examinees as well as their perceived authenticity in requiring examinees to integrate substantive content from reading and listening activities in a manner that resembles ordinary academic writing. Because these integrated tasks call for examinees to use certain rhetorical, ideational, and stylistic features, they may prompt ESL/EFL assessors who judge them to balance their attention evenly among the ideas, organization, and language features of ESL/EFL students' writing.

But this very quality of the prototype writing tasks also requires that they be designed with the utmost care and diligence. They must specify clearly and precisely, both for examinees in their writing and for raters in their assessments, how each particular task is expected to make use of source materials from companion reading or listening materials, the conditions under which the writing is to be performed, the criteria by which effectiveness on the task is to be evaluated, and differences between the purposes and difficulties of the respective tasks. A further consideration, not accounted for in the design of the present research, is that the participating assessors all read compositions on paper, scanning them routinely for overall layout on the page, whereas raters in ETS's Online Scoring Network read compositions on a computer, screen by screen, typically without being able to perceive the overall layout of a text at a glance. Plans for the TOEFL 2000 examination will need to consider in advance whether both handwriting and typing will be permitted in the test and what the interface for scoring compositions will be, because this may well affect raters' decision making or the kinds of training they might require.

References

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

Cohen, A. (1994). Verbal reports on learning strategies. In A. Cumming (Ed.), Alternatives in TESOL research: Descriptive, interpretive, and ideological orientations. *TESOL Quarterly, 28*, 678-682.

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

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

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

Cumming, A. (1997). The testing of second-language writing. In D. Corson (Ed.), *Language assessment, Vol. 7. The encyclopedia of language and education* (pp. 51-63). Dordrecht, Netherlands: Kluwer.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series Report No. 18). Princeton, NJ: Educational Testing Service.

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

Erdosy, 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, Toronto, Canada.

Ericsson, K., & Simon, H. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Freedman, S., & Calfee, R. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. Walmsley (Eds.), *Research on writing* (pp. 75-98). New York: Longman.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second-language writing in academic contexts* (pp. 241-278). Norwood, NJ: Ablex.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*, 337-373.

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

Henning, G. (1991). Issues in evaluating and maintaining an ESL writing assessment program. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 279-292). Norwood, NJ: Ablex.

Huot, B. (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 students' essays. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206-236). Cresskill, NJ: Hampton Press.

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

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 abilities. *Annual Review of Applied Linguistics, 18*, 219-240.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33-69). Urbana, IL: National Council of Teachers of English.

Mendelsohn, D., & Cumming, A. (1987). Professors' ratings of language use and rhetorical organization in ESL compositions. *TESL Canada Journal, 5*, 9-26.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3: Performance testing, cognition and assessment* (pp. 92-111). Cambridge: Cambridge University Press.

Minitab, Inc. (1997). MINITAB™ Statistical Software (Release 12) [Computer software]. State College, PA: Author.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*, 101-143.

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

Purves, A. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English, 26*, 109-123.

Raimes, A. (1990). The TOEFL Test of Written English: Causes for concern. *TESOL Quarterly, 24*, 427-442.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: A study of how raters evaluate ESL compositions on a holistic scale. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129-152). Cambridge: Cambridge University Press.

Santos, T., & Kantor, R. (n.d.). *First-stage operationalization of writing variables in TOEFL 2000: Task prototyping and piloting*. Unpublished manuscript.

Smagorinsky, P. (Ed.) (1994). *Speaking about writing: Reflections on research methodology*. Thousand Oaks, CA: Sage.

Song, B., & Caruso, L. (1996). Do English and ESL faculty differ in evaluating the essays of native-English speaking and ESL students? *Journal of Second Language Writing, 5*, 163-182.

Stansfield, C., & Ross, J. (1988). A long-term research agenda for the Test of Written English. *Language Testing, 5*, 160-186.

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. 111-126). Norwood, NJ: Ablex.

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

Wolfe, E., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*, 465-492.

Appendix A: Profile Questionnaire

The purpose of this questionnaire is to gather background information related to the data you generated in the think-aloud protocols while you assessed the ESL compositions for the project, *An Investigation into Raters' Decision Making, and Development of a Preliminary Analytic Framework, for Scoring TOEFL Essays and TOEFL 2000 Prototype Tasks*. Please note that the aim of the research is not to judge your performance, but, rather, to understand it more fully. As with other data generated by the project, your identity will remain confidential.

Please complete the questionnaire shortly after assessing the papers assigned to you. This will, on the one hand, prevent the questionnaire from influencing your rating process and, on the other hand, should allow you to recall clearly what you did in the process of generating the think-aloud protocols.

Your name (or pseudonym) is: _____

Date of completing this questionnaire: _____

I. Your Assessments

1. Reflecting back on the assessments you have just done for this project, to what extent did any particular assessment scheme(s) (e.g. rating scales, checklists, research instruments, etc.) influence you in assessing the compositions? Please circle the number that best corresponds to your answer.

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

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

_____

_____

_____

3. Reflecting back on the assessments you have just done for this project, what were the three most important factors in your past experiences (i.e., work as an assessor, instructor, researchers, etc.) that you think may have influenced your assessment of the compositions? Please describe these briefly:

i) _____

ii) _____

iii) _____

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. Your gender is: Male _____ Female _____

6. Your age is: < 30 ___ 31-40 ___ 41-50 ___ > 50 ___

III. Current Professional Status

7. Your current role(s) is/are: Assessor ____ Teacher ____ Administrator ____ Student ____ Researcher ____ Other (specify)_____

8. The context(s) you have mostly worked in is/are: English (mother tongue) ____ ESL ____ EFL ____ ESP ____ Other (specify) _____

IV. Language(s)

9. Your first language is: _____

10. Your dominant language at home at present is: _____

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

V. Educational History

Please describe your educational background in terms of:

	Degree/Diploma/ Certificate	Subject Area	Language of Education
12.	Undergraduate Studies:		
	_____	_____	_____
13.	Postgraduate Studies:		
	_____	_____	_____
	_____	_____	_____
14.	Professional Certification:		
	_____	_____	_____
15.	Any Specialized Training Related to Assessment:		
	_____	_____	_____

VI. Professional Writing Experience

Please characterize, in two or three brief statements, your professional experience in the following areas. Indicate publications and routine writing in work contexts, 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 and/or Assessing Writing

19. Please list under the following headings your three most significant teaching and/or assessment experiences, such as might have influenced your rating of ESL compositions in the present research project:

Institutional Context	Language(s)	Number of Years You Did This
_____	_____	_____
_____	_____	_____
_____	_____	_____

20. How would you describe your own skill in assessing ESL writing?

_____ Expert _____ Competent _____ Novice

21. How many years' experience do you have in assessing ESL writing?

< 2 ____ 3 to 4 ____ 5 to 6 ____ 7 to 10 ____ > 10 ____

22. Have you given any training courses in assessing language performance or administered such programs? If so, please describe these briefly.

_____

_____

_____

_____

Thank you for this information!

Appendix B: Instructions for Think-Aloud Protocols

Please read these instructions through carefully before you begin the assessments.

Purpose

These instructions are written to help guide you, and others producing think-aloud protocols for this project, in a consistent and informative manner. Think-aloud protocols ask people to say everything they think about while they perform a task, with the aim of documenting and better understanding what you pay attention to and consider important when you do a task. The purpose of the think-aloud protocols for this study is to find out in as much detail as possible what you, as an experienced assessor of ESL compositions, are thinking about, deciding, and doing while you rate a sample of ESL compositions. The most important thing to emphasize is, say everything you are thinking about, and make certain this is recorded clearly onto the tape recorder. What you say will become important data for our research. Thanks!

The Assessment Task

You will receive a package of 60 written compositions produced by adult ESL learners of varying levels of proficiency in English. You are asked to assign a number from 1 to 6 to each composition based on your assessment of the quality of each composition (i.e., 6 = most proficient writer in English, 1 = least proficient writer in English, with numbers 5, 4, 3, and 2 in between these extremes). Say as much as you can while you are reading the composition and deciding on how to rate it, and be sure the number you assign to each composition is recorded along with your ongoing impressions of it.

The Compositions

You will also receive copies of the essay prompts originally given as instructions to the students who wrote the compositions, so you know what they were asked to do and write about. There are four different essay topics. Please keep these topics fully confidential and secure, because they are from a real examination context. These compositions have been written under exam conditions within 30 minutes. Some are written by pen and paper and some on a word processor. The authors of the compositions come from diverse parts of the world; as you'll see, their abilities in writing vary a lot. The compositions have been identified with code numbers, but we have tried to arrange the codes so that you do not know the original score given to any composition (nor should you try to guess this). The order in which you receive the compositions has been randomly sequenced, but you should receive together about 10 to 15 compositions on each of the four different topics. The actual number varies for each person, because we don't want you to be guessing how many scores to give to the compositions.

The Ratings

In making your assessment, try to avoid considering any existing rating scales as a basis for your judgments, because the purpose of this research is not to try to validate any existing rate scale, but rather to create a new one, based on analyses of assessors' decision-making processes. However, if you do think of a particular rating scale (or even several of them) that you are familiar with, by all means speak about this onto the tape recorder. Such data are important to us. The rating will not be judged as right or wrong, nor will we be analyzing the final numbers you assign in these ratings. Our interest is in the spoken data on what you think about while you make the ratings. We have intentionally not told you what criteria to use to rate compositions as either 1, 2, 3, 4, 5, or 6 because we want to find out the criteria that you, independently, use.

Recording Your Thoughts While Assessing

* Keep talking, conveying your thoughts continuously, while you assess the compositions, from the initial point when you first see each composition until you have completed rating it, and indeed until you rate the whole set of them.

* Speak in English as much as you can. If you use another language, please translate it onto the tape (so we can understand this later in our analyses).

* Speak continuously. Report fully, even what might seem trivial. Do not assume that others know what you are doing or thinking.

* Try to avoid speech fillers (i.e., uh, um) as much as possible. Try to use words instead, so that we can understand what your thoughts have been.

* Talk and make your assessment as naturally and as honestly as you can, according to what you usually do when you assess students' compositions. Don't start rationalizing your ideas at length; we're just interested in your decisions as you make them, and as you would do if you were doing this task on your own.

Instructions for Taping

1. Set up the tape recorder and check that it works. Check whether it records properly and that the quality of the tape is okay by trying out a few words initially, then playing it back. Make sure there is no background noise (e.g., fans, music, foot tapping, etc.).

2. Set the microphone at an appropriate distance close to your face and voice, and try not to block the microphone with papers or your hands. Be sure that the VOR (voice-activated-recording) and Pause switches are off!

3. Turn the tape recorder on, and record the date and your pseudonym to be used in the research.

4. Open the package and begin to assess the compositions. As you assess each composition, indicate clearly the code number of the composition that you are rating when you start to read it. Then when you have made a rating decision, indicate the score (from 1 to 6) you assign to it. Then indicate the code number of the next one as you start to assess it, and so on.

5. Report your first impression of each composition and if it influences your rating. Then continue talking, saying what you are thinking about, as you are making your assessment decisions.

6. It is not necessary for you to write anything for this task. However, if you write something down (i.e., marks, comments, corrections), say that you are writing, and report what you are writing. Feel free to write on the compositions, if you like, but we will not be analyzing any written notes you make (so we want to know on the tape recording what you might write).

7. You may read the compositions aloud or silently, according to what feels most "natural" to you. Make sure you report exactly what you are doing. If you are reading silently, indicate which part of the composition elicits your comments.

8. If you happen to reconsider any of your ratings (e.g., for a second or third time), verbalize your reason(s) for doing so, and indicate on the tape that this is what you are doing.

9. You may use more than 1 tape, or even 2 or 3 of them. As you finish each tape, label it with your pseudonym, the sequence of the tapes (e.g., tape 1 out of 2), and the date. On the tape, as well, record at the start of the tape, whether it is the first, second, or third tape you are using.

10. If you have to take a break while you are assessing the compositions, indicate on the tape that you are doing this, turn the tape recorder off, then when you start again, indicate this clearly onto the tape.

11. When you have completed assessing all of the compositions in the package, then indicate this clearly on the tape, so we know when your think-aloud protocol ends.

12. At the end of the assessment session and tape recording, please put all the compositions, together with the tape(s) you have used, back into the package. Thanks!

Appendix C: Transcription Conventions

General Layout

* Identify transcripts at the top left corner of each page using running heads that indicate the person's pseudonym and date.

* Number all pages in the top right corner of each page.

* Set off each person's consideration of each new composition, using a line of four asterisks (e.g., ****).

* Similarly set off any procedural behaviors (e.g., sorting essays, reflections on rating behavior) with a line of four asterisks.

* Use 1.5 lineation throughout.

* Use the font New Times Roman 12 to print the transcripts.

* Print only on one side of the paper.

Symbols

( )	for uncertain transcription
x	incomprehensible item, one word only
xx	incomprehensible item of phrase length
xxx	incomprehensible item beyond phrase length
...	three dots indicate a pause of five seconds or more
-	a hyphen indicates an incomplete word (e.g., wait plea-)
—	three hyphens indicate an interrupted or incomplete sentence
[ ]	square brackets indicate a word or phrase in a language other than English, naming the language as well (e.g., [French])
" "	quotation marks indicate text read directly from the original composition

"USE CAPITAL LETTERS ..." for long stretches of text read aloud. Use quotation marks and indicate only the first three words of the text read in capital letters (i.e., it is not necessary to transcribe verbatim whole passages that are read aloud, but it would be helpful to identify them, if we want to look back at them later).

For standard hesitation markers or exclamations, use:

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

Note any observations or interpretive comments at the end of the transcripts (not in the middle).

Appendix D: Decision-Making Behaviors While Rating ESL Compositions[2]

Self-monitoring focus	Task fulfillment: rhetorical and ideational focus	Language focus
Interpretation strategies		
* scan whole text * envision situation of writer * focus self on task rubric	* interpret ambiguous phrases * discern rhetorical structure * summarize propositions	* classify error types * "edit" phrases for interpretation
Judgment strategies		
* establish personal response * define and revise own criteria * compare with other compositions or anchor papers * distinguish interactions between categories * summarize judgments collectively * articulate scoring decision	* assess total output * assess relevance * assess coherence * assess interest * identify redundancies * assess topic development * assess helpfulness in guiding reader * rate content and organization overall	* establish level of comprehensibility * establish error values * establish error frequency * establish command of lexis * establish command of syntax and morphology * establish command of spelling and punctuation * rate language overall

2 Initial version of a descriptive framework, as presented in the proposal for this project, derived from Cumming, 1990, and Sakyi, 1997.

Appendix E: Examples of Phase One Decision-Making Behaviors

Self-Monitoring Focus – Interpretation Behaviors

IS1. Read or interpret essay prompt.

"Uh, one of the things I'm doing right now is to uh, check back uh, on the uh, topic because uh..." (Scott)

"Let me see the topic again... the prompt is do you agree or disagree, give specific reasons and examples to support." (Patricia)

IS2. Read or reread composition.

"Um ... I'm rereading this again silently. I think I'd better read this in silence." (Melissa)

"Just looking back at it, reading through it again..." (Jane)

IS3. Envision personal situation of the writer.

"Uh, so this person went back and read over what he or she had written..." (Patricia)

"... is this person ready for university admission? I suppose this person could do an undergraduate degree but they probably would not get very high marks in, uh, their writing." (Roy)

IS4. Scan whole composition.

"Just looking over it to look at the length. It completely fills the given writing space." (Jane)

"I am scanning over the writing again, to see if I can find any redeeming features." (Jane)

Self-Monitoring Focus – Judgment Behaviors

JS1. Decide on macrostrategy for reading and rating.

"Actually, I should probably go over my 2s and 3s again before I finish marking." (Zoey)

"I think I will just take the compositions in the order in which they come uh, because uh, I don't really feel like reordering them." (Scott)

JS2. Consider own personal response or biases.

"I'm sort of disagreeing with the ideas. I, I think that uh, I try not to uh, penalize students because they provide ideas which I consider to be stupid..." (Scott)

"Maybe I'm being hard on this person because my subjective response to the essay is positive and I don't want that to influence me." (Zoey)

JS3. Define or revise own criteria.

"The student has very repetitive grammar structure in the sentences, which to me, indicate that this student is a kind a beginning level writer." (Paul)

"So my scale's, is changing now." (Zoey)

JS4. Compare with other compositions or "anchors."

"Uh, but uh let me compare it back to some of the others I've read here." (Roy)

"It's not as good as some of the 4s." (Zoey)

JS5. Summarize, distinguish, or tally judgments collectively.

"There are some problems, some grammatical mistakes, but as I said they're not, uh, they don't interfere with comprehension, or even the, you know, the comprehension of each sentence." (Karen)

"Again, there's no essay, academic essay structure imposed in this essay. The sentences are repetitive, rather simplistic. The vocabulary is simplistic. There's no thesis statement." (Paul)

JS6. Articulate general impression.

"It's very weak." (Gary)

"Uh, that's low." (Zoey)

JS7. Articulate or revise scoring

"I give 3 to this composition." (Ed)

"I think it should maybe be a 4, rather than a 5. Let me change that to a 4." (Roy)

Rhetorical and Ideational Focus – Interpretation Behaviors

IR1. Interpret ambiguous or unclear phrases.

"I know what the student uh, tries to say which is that 'I uh, not only do I like to study literature.'" (Scott)

"What does it mean, 'GENERAL'? 'THE PEOPLE WHO ...' You mean people are easy to understand or these programs are easy to understand?" (Gary)

IR2. Discern rhetorical structure.

"It is a, a typical essay format where they've stated the, stated the intention, given a topic sentence and then summarized at the end ..." (Zoey)

"There is an introduction, there is a one sentence conclusion, uh, and three arguments ..." (Patricia)

IR3. Summarize ideas or propositions.

"So the student uh talks about the room and its condition in his hometown, and these are extremely crowded, and people live in simple style flats." (Ed)

"So, this person is proposing a long vacation each year rather than having several short vacations." (Gary)

Rhetorical and Ideational Focus – Judgment Behaviors

JR1. Assess reasoning, logic, or topic development.

"The writer did not elaborate on the thesis sentence, the topic ..." (Ed)

"It's a lot of assumptions here that hasn't, haven't been defended." (Gary)

JR2. Assess task completion.

"O.K., the composition is uh not finished, uh it's not complete." (Ed)

"He is missing the task." (Gary)

JR3. Assess relevance.

"This is absolutely irrelevant and off topic sentence." (Patricia)

"... although the one sentence answers the question, none of the others do ..." (Jane)

JR4. Assess coherence.

"I think the argument is somewhat convoluted ..." (Zoey)

"In terms of coherence, there's certainly a focus." (Karen)

JR5. Assess interest, originality, or creativity.

"The reason I didn't like it was because I found the ideas boring …" (Zoey)

"... just looking back at the beginning, uh, the idea of approaching the topic by pretending that a discussion had taken place, I like that, it's creative." (Jane)

JR6. Identify redundancies.

"The second sentence is just a repetition of the first ..." (Patricia)

"Uh, the same arguments are brought up a few times, so repetition of ideas." (Karen)

JR7. Assess text organization.

"Basically, the whole thing guides the reader quite well, finishes off with a conclusion." (Karen)

"... the composition is not organized in paragraphs. And ... it's very difficult for me to read and follow." (Ed)

JR8. Assess style, register, or genre.

"It's got this bookish version of English, formulaic way of writing compositions." (Roy)

"It says 'IMAGINE THIS' at the beginning of the second paragraph and this style is not, to me, it does not look appropriate to the situation of this writing." (Melissa)

JR9. Rate ideas or rhetoric.

"... the first reason is not very illuminating, the second one is." (Scott)

"It's not rhetorically very sophisticated." (Roy)

Language Focus – Interpretation Behaviors

IL1. Observe layout.

"... this is a typewritten one. It looks neat ..." (Melissa)

"... uh, this is terrible handwriting." (Gary)

IL2. Classify errors into types.

"So, the, the person is definitely having problems with the article, with prepositions, with uh, agreement in some cases, subject/verb agreement." (Zoey)

"So the errors uh, again tend to occur more in the areas of uh, prepositions, occasionally article usage uh, sometimes uh, overuse of adjectives like 'good' and 'big'." (Scott)

IL3. Edit phrases for interpretation.

"I think they are trying to say 'within' 'TEN YEAR'; looks like a typing mistake here." (Jane)

"I think this should be just 'unwind'..." (Roy)

Language Focus – Judgment Behaviors

JL1. Assess quantity of total written production.

"Uh, three short paragraphs, handwritten." (Scott)

"... I see seven paragraphs." (Ed)

JL2. Assess comprehensibility.

"The last paragraph seems to be a little uh, confusing ..." (Gary)

"It's not comprehensible. I don't know what the person is trying to say." (Roy)

JL3. Consider gravity of errors.

"Some misspelling although they are not important or most of the misspellings are for minor words like 'AND' and 'AN' or 'AND'." (Paul)

"Oh, some dramatic grammatical errors, so it affects the coherence of the composition." (Ed)

JL4. Consider error frequency.

"About every second word is inaccurate, or limited, or wrong, or confusing." (Roy)

"Very few errors of any kind uh, occasional spelling errors uh, which may even be typos uh, uh, but almost no, no problems whatsoever." (Scott)

JL5. Assess fluency.

"So, the person has good ideas but they're not always uh, clearly and fluently expressed..." (Zoey)

"Almost like he has the fluency of a native speaker, but he has little control over punctuation..." (Gary)

JL6. Consider lexis.

"This student's vocabulary seems to be rather simple." (Paul)

"O.K. the vocabulary is pretty good, words like 'CREDIBLE', 'TOTALITY', 'CONCRETE'." (Jane)

JL7. Consider syntax or morphology.

"'IT IS IMPORTANT...' Huhh, this is really limited syntax." (Roy)

"He uses transition words, but the sentence grammar is awful." (Paul)

JL8. Consider spelling or punctuation.

"... so many spelling errors that they are distracting." (Scott)

"... he's got word comma, a word, comma, a word, comma, and etc., uh which he shouldn't be doing." (Paul)

JL9. Rate language overall.

“This person has a good control of language.” (Gary)

“... uh language use is not strong.” (Melissa)

Test of English as a Foreign Language
P.O. 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: http://www.toefl.org

57906-010266 • Y121M.75 • Printed in U.S.A.
I.N. 990630