TOEFL®

Monograph Series

MS - 18
APRIL 2000

TOEFL 2000 Writing Framework: A Working Paper

Alister Cumming
Robert Kantor
Donald Powers
Terry Santos
Carol Taylor

ETS™ Educational Testing Service

TOEFL 2000 Writing Framework: A Working Paper

Alister Cumming
Robert Kantor
Donald Powers
Terry Santos
Carol Taylor

Educational Testing Service
Princeton, New Jersey
RM-00-5

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

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

E-mail: toefl@ets.org

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

Foreword

The TOEFL Monograph Series features commissioned papers and reports for TOEFL 2000 and other 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 the new forms of assessments 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 are the working papers that lay out the TOEFL 2000 conceptual frameworks with their accompanying research agendas. The initial framework document, Monograph 16, describes the process by which the project will move from identifying the test domain to building an empirically based interpretation of test scores. The subsequent framework documents, Monographs 17-20, extend the conceptual frameworks to the domains of reading, writing, listening, and speaking (both as independent and interdependent domains). As such, the current frameworks do not yet represent a final test model. The final test design will be refined through an iterative process of prototyping and research as the TOEFL 2000 project proceeds.

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 paper builds on *TOEFL 2000 Framework: A Working Paper* (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 1999) by setting out a preliminary working framework for the development of the writing assessment component of the TOEFL 2000 test.

The monograph is organized into four major parts. Following a brief introduction, Part 2 presents a conception of writing proficiency and focuses in particular on academic writing—the domain of the TOEFL 2000 test. The third part presents an initial writing framework for the test; it reviews the test domain, proposes an organizational scheme, and identifies the task variables of interest. The fourth section lays out an initial research agenda for establishing the validity of interpretations, and the appropriateness of action, that would result from the introduction of writing measures growing out of this framework and approach to test design. The paper concludes with a discussion of the important ways in which the TOEFL 2000 approach to testing writing is intended to improve on its predecessors.

Key words: communicative competence, writing proficiency, writing for academic purposes, reader-writer model, scoring methods

Acknowledgments

The authors gratefully acknowledge the support, comments, and suggestions of our many colleagues, in particular those on the TOEFL Committee of Examiners, on the ETS staff, and on the TOEFL 2000 teams working on the speaking, listening, and reading frameworks.

Table of Contents

I. Introduction

This paper builds on *TOEFL 2000: A Working Paper* (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 1999) by operationalizing a working framework for the development of the writing assessment component of the TOEFL 2000 test. The Jamieson et al. paper presents a rationale for conceptualizing a test framework and lays out a sequence of six steps from identifying and organizing the test domain, to identifying task characteristics and operationalizing their associated variables, to providing an empirically based interpretation of test scores.

The framework set out in this paper is preliminary; it provides an initial working definition of the TOEFL 2000 writing domain, a set of writing task characteristics and variables believed to account for variance in test performance, an accompanying research agenda, and criteria for a better test. Conceptually, we have followed the TOEFL 2000 framework but have deviated slightly from certain categories outlined in it. Some categories, such as those included in *Situation* (Jamieson et al., 1999, pp. 14-16), are less salient to the nature of the writing tasks and types of writing that are relevant to the TOEFL 2000 test. We expect that the final writing framework will be shaped and refined through research on the TOEFL 2000 writing domain as well as consideration of the other TOEFL 2000 skill-area frameworks and research (i.e., reading, listening, and speaking frameworks).

The following section of this document, Part 2, presents a conception of writing proficiency and focuses in particular on academic writing – the domain of the TOEFL 2000 test. The third part presents the writing framework for the test; it reviews the test domain, proposes an organizational scheme, and identifies the task variables of interest. Part 4 reviews the research agenda for the new writing test, and the fifth and final section discusses the important ways in which the TOEFL 2000 writing test will improve on its predecessors.

2. Conceptualizing Writing Proficiency

> *Logically and empirically, there seems little merit in references to a unitary construct [of] writing* (Purves, 1992).
>
> *Agreeing on a global concept of writing is something that is difficult to imagine university faculty as being able to do* (White, 1995).

These quotes suggest that developing an entirely adequate statement of the target construct for a test of writing, much less second-language writing, will at best be difficult. A reasonable first step, we suggest, is to set a modest goal: to aim for a "conception" of writing proficiency rather than for a more scientifically rigorous "construct." Considering various facets of writing proficiency from a variety of perspectives, we hope to converge on a workable conception that can guide further research and development.

A Broad View

Writing is, according to Hamp-Lyons and Kroll (1997), "an act that takes place *within a context*, that accomplishes a *particular purpose*, and that is appropriately shaped for its *intended audience*" (p. 8). It is, by nature, a *communicative* act: "We write mainly to communicate with other humans" (Hayes, 1996, p. 5). Although some (Barzun & Graff, 1977) believe that there is no essential difference in the form and the processes that are employed in the creation of any document – whether a student book report, a CEO's report to stockholders, or the President's State of the Union address – we will find it useful here to distinguish specific purposes for writing as well as different genres of writing.

Language assessment prior to and since Carroll's (e.g., 1975) widely cited skills model has conventionally defined writing as one of four basic, integrated language skills. Writing differs from the other three skills (i.e., reading, listening, and speaking) in many obvious but important ways. For instance, writing is primarily a productive endeavor, involving the utilization of numerous subcomponents and unique types of skill and knowledge. At the same time, writing utilizes many of the linguistic resources common to the other skills, such as speaking, though in the form of heavily conventionalized scripts and text forms. When employed in writing, these scripts and text forms may require extensive reflection and cognitive effort (Bereiter & Scardamalia, 1987) so that, when read, the written texts can communicate these thoughts independently, without the presence of a direct interlocutor (Olson, 1996). Sperling (1996) has provided a thoughtful review of the relationships between speaking and writing, summarizing both the ways in which speaking may facilitate writing as well as the ways in which it may hinder it.

Regardless of the precise relationships among language modalities, it is clear that for communication to occur, simply knowing the conventions of standard written English, as important and fundamental as this may be, is insufficient. The essence of writing is the generation of ideas and the translation of these ideas to the printed word. This fundamental notion has been approached from several different traditions (Hamp-Lyons & Kroll, 1997): writing has been viewed alternatively as

- the transmittal of knowledge,
- the creation of knowledge, and

- a channel for the human imagination.

Writing can take many forms: lists, outlines, sketches, proposals, drafts, comments, etc. It depends to some degree on how much "independent invention" is entailed. Real-life writing seldom calls for the writer to start completely from scratch, but rather to draw on a variety of resources that are available during the writing process – previous documents, genre models, and other sources, for example (Van der Geest, 1996), and, of course, the writer's own prior knowledge and experiences.

Writing can be viewed in terms of its discourse functions (e.g., argumentative, narrative, and reflective), its cognitive demands, and its social nature (Purves, 1992). It is, to be sure, an intellectual activity that requires cognitive processes (such as long-term and working memory), but it also has an affective side that involves motivation (Hayes, 1996).

Various cognitive models have been advanced to show that, although one can focus on the product that results from writing, writing clearly entails a *process*. Modern conceptions suggest that this process is not easily described in terms of neat, sequential, additive steps. Rather the writing process is recursive, iterative, and nonlinear in nature. Writing is additive only in the sense that appropriately chosen words contribute to meaningful sentences, which in turn constitute increasingly large segments of meaning (e.g., paragraphs, sections, chapters, volumes).

A recent cognitive model, as set forth by Hayes (1996), includes two major components: the task environment and the individual. The former includes a social component (the audience, the environment, and other sources that the writer may access while writing) and a physical component (the text that the writer has written so far and the medium used to produce it, e.g., a word processor). The individual component includes four distinct subcomponents:

- motivation and affect,
- cognitive processes (text interpretation, reflection, and text production),
- working memory, and
- long-term memory (including knowledge of topics, genres, audience, and task schemas, e.g., for editing).

Planning has a place in all of the cognitive models. The implication is that most writing requires thought and reflection rather than impromptu reactions.

Writing is largely a *social* enterprise. For instance, it sometimes entails a collaboration among several writers. But the major social aspect of writing is its relationship to an audience of readers. Although writers may sometimes see themselves as the primary consumers of their writing (as in note-taking, keeping a diary, or writing to learn), most writing is undertaken with some other, external audience in mind – be it a professor assessing a student's understanding of a subject, a shareholder reviewing a company's annual report, or a trained essay reader evaluating (from an examinee's test essay) an applicant's readiness for further education. Indeed, the audience/reader plays a central role in determining what constitutes good writing, for the ultimate test of a writer's skill is

- the extent to which readers are informed, persuaded, taught, directed, etc., and

- the degree to which readers themselves are satisfied that these functions have been accomplished without their having had to exert an inordinate amount of effort.

Barzun and Graff (1977) believe that the writer's task is especially difficult because of a duty to inform several audiences – not only colleagues, for instance, but also the "Unknown Reader," whom a writer can never fully anticipate. As Sperling (1996) notes, writing entails, to a large degree, anticipating how one's words will be read and understood.

Responsibility for conveying meaning does not fall solely on the writer, however: readers themselves bear some responsibility to be attentive and open-minded, for example, while retaining a healthy skepticism about the writer's claims. What can be considered to be well written (or easily readable) is to some extent a function of the characteristics of readers – for example, the extent to which they share the same backgrounds, knowledge, and experiences as the writer.

One predominant view of writing is that it involves the production of an *extended* piece of writing. Here, the concepts of elaboration, conciseness, and redundancy all come into play. Elaboration may be necessary to convey meaning, and redundancy may be required at times to ensure understanding. Furthermore, these qualities vary extensively with the genres or types of writing being considered, with the purposes and context of writing, and with the specific communities judging their qualities. The challenge to the writer, typically, is to determine how much elaboration is needed to inform (or persuade, etc.) and engage an audience without being tedious. Here, too, the reader may have some responsibility to reveal his/her expectations to the writer – for an idea, an argument, etc., can often be explained with varying degrees of comprehensiveness, for example, in a volume, a chapter, a page, or "25 words or less." Appendix A outlines in greater detail various standards generally considered important for judging the quality of expository writing.

Although the preceding discussion has helped, we think, to converge on a general conception of writing, we will need to limit our focus further, to writing as it is practiced in an academic setting. Next, therefore, we focus more specifically on writing in an academic context.

Writing for Academic Purposes

The focus of the TOEFL 2000 test is on writing in academic settings. For academic as well as for other purposes, written texts are produced both *in* and *for* specific social contexts, involving communities of people who establish particular expectations for genres and standards for writing within these communities, both locally and universally. In theory, writing in academic settings typically involves the production of written text forms that, by virtue of the expectations of academic institutions and the interactions among members of the subcultures they represent, eventually become conventions (Berkenkotter & Huckin, 1995; Connor, 1996; Paltridge, 1994; Swales, 1990). Students who enter colleges or universities necessarily require a core set of these genres for academic writing in order to complete course-related assignments and to display the knowledge they are acquiring (e.g., Hale, Taylor, Bridgeman, Carson, Kroll, & Kantor, 1996; Waters, 1996; Weir, 1984). Many of these genres are ones that students develop as undergraduates (or continue to develop as graduate students or on the job); so, one immediate issue for research will be whether it is reasonable to expect students to have complete

knowledge or mastery of specific text genres when they begin undergraduate study. In many undergraduate contexts, the emphasis of writing is often on telling people about the knowledge one has, or is acquiring, rather than on using writing to create unique or novel knowledge, as might be expected in graduate studies (Collis & Biggs, 1983; Hale et al., 1996). Langer and Applebee (1987) further describe the functions of writing in academic classrooms "to draw on relevant knowledge and experience in preparation for new activities; to consolidate and review new information and experiences; to reformulate and extend knowledge" (p. 41). Focusing on transmitting, rather than creating knowledge, is consistent with our primary interest: individuals' writing and language abilities, rather than their academic knowledge or expressive creativity per se.

A central concern in designing the writing assessment for the TOEFL 2000 test will be to select tasks that represent key genres for writing which are integral to as wide a range of university or college contexts as possible, without biasing this selection in favor of (or against) particular groups, areas of interest or knowledge, or specific situations. However, as Paltridge (1994) notes, although conceptually useful, definitions of genres in writing are extraordinarily varied and difficult to make general claims about. Rather than using the Jamieson et al. framework, organized around the headings of situation, text, and rubric, we have decided that our conceptualization of academic writing can more effectively be presented in terms of the headings of task stimuli, rhetorical functions, topic characteristics, and evaluative criteria.

Assessment of Academic Writing

As documented elsewhere (Breland, 1996; Grabe & Kaplan, 1996; Hamp-Lyons, 1990; White, 1988, 1995), the domain of academic writing in many large-scale writing assessments has moved away from traditional multiple-choice assessments toward the use of more performance-based assessments. A few examples of large-scale assessments that have incorporated direct measures of writing include the National Assessment of Educational Progress (NAEP), the International Association for the Evaluation of Educational Achievement (IEA), the Graduate Management Admission Test (GMAT®), the Graduate Record Examination (GRE®), the TOEFL Test of Written English (TWE®), and the International English Language Testing System (IELTS). The writing assessments included in these tests reflect a range of writing purposes and tasks.

The framework for the 1998 NAEP writing assessment was developed using a national consensus process which revised some aspects of the 1992 framework but retained three primary purposes for writing among students in grades 4, 8, and 12: narrative, informative, and persuasive (Applebee, Langer, Mullis, Latham, & Gentile, 1994). The 1998 framework specified a number of overarching objectives (The National Assessment Governing Board, n.d.), stating that students should write for a variety of purposes, on a variety of tasks, from a variety of stimulus materials; generate, draft, revise, and edit their writing; display effective choices in the organization and elaboration of their writing; and value writing as a communicative activity (p. 5).

Like NAEP, the IEA study of written composition assessed writing in primary and secondary school settings. However, the IEA study examined writing in the schools of 14 countries. The IEA writing framework (Gorman, Purves, & Degenhart, 1988) drew largely on the work of Baker and Quellmalz (1978) and Baker (1982) to create a writing domain specification that included four types of writing

tasks: (a) pragmatic, summary, and descriptive tasks, (b) personal narrative tasks, (c) persuasive tasks, and (d) reflective tasks (Gorman et al., 1988, p. 37).

Among writing assessments used for admissions purposes in higher education, the GMAT writing assessment comprises two 30-minute writing tasks where the domain of writing is defined as an "analysis of an issue" and an "analysis of an argument." Issues and arguments are often presented as quotations, and the tasks are intended to measure examinees' ability to think critically and communicate ideas in writing (Breland, 1996). The GRE testing program has announced plans to introduce at least one 45-minute, expository essay in 1999 (Breland, 1996). The writing task will likely require examinees to take a position on a complex issue and support that position with reasons and examples drawn from their academic studies and/or personal observations and experiences (Powers, Fowles, & Boyles, 1996).

The TOEFL TWE test (ETS, 1996) uses brief, simply stated expository writing tasks that are "designed to give examinees the opportunity to develop and organize ideas and to express those ideas in lexically and syntactically appropriate English" (p. 7). In most cases the writing domain of TWE questions asks examinees to take a position and support it with reasons and/or examples (p. 25). Another second language writing test, IELTS, requires examinees to write on two prompts but allows them to choose between an "academic writing module" or "general training writing module" (IELTS Handbook, 1995; Alderson, 1993).

3. Writing Framework for the TOEFL 2000 Test

The TOEFL 2000 writing domain must, as a starting point, be consistent with the conception of writing outlined in the previous section. It must also be an outgrowth of the domain identification established for the TOEFL 2000 project, that is, to measure examinees' English-language proficiency in situations and tasks reflective of university life in North America and to be used as one criterion in decision-making for undergraduate and graduate admissions. In addition to being constrained by the TOEFL 2000 test purpose, the writing framework must work within the practical constraints of the testing situation (e.g., time, cost, security, available technology). The TOEFL 2000 working framework (Jamieson et al., 1999) organizes the test domain by modality (i.e., reading, listening, speaking, and writing), but notes that these modalities can be tested both independently and integratively and, in fact, envisions a number of test tasks that will assess the skills integratively.

Our task is one of evaluating the extent to which our measure (a) faithfully represents important facets of this conception and (b) does not inadvertently measure traits, tendencies, styles, etc., that are extraneous to our conception. No single measure administered in a limited period of time can be expected to capture all of the important aspects of writing proficiency, and no assessment will be entirely free from every irrelevant influence. The objective, therefore, will be to mount research and development activities that will enable us to evaluate the tradeoffs as we attempt to maximize construct representation and minimize construct-irrelevant variation in test scores.

Identifying the Test Domain

Considering the purposes of the TOEFL 2000 test, the domain of the writing framework is organized around eight overarching objectives or working assumptions. Namely:

1. *The TOEFL 2000 writing assessment should require students to produce and sustain in writing coherent, appropriate, and purposeful texts in response to assigned tasks.*

In order to be credible, the writing assessment must include at least one task where examinees have an opportunity to produce an elaborated written response of sufficient length to demonstrate clear writing competence on rhetorical, lexical, and syntactic levels. A minimum of 30 to 40 minutes is viewed as obligatory in providing adequate time for writers to generate, draft, and edit written ideas for an elaborated response (Hale, 1992; Hamp-Lyons & Kroll, 1997; Powers & Fowles, 1996; Powers & Fowles, 1998). This writing sample may be in response to either an independent invention or interdependent text-based prompt (see discussion later in this section).

2. *The TOEFL 2000 writing assessment should allow for a total writing time of between 60 to 75 minutes.*

Testing time is among the practical constraints that must be weighed in developing the next generation of a computer-based TOEFL test that will be delivered on demand to almost a million individuals each year. We assume a maximum total testing time of 4.5 hours for the TOEFL 2000 assessment of reading, listening, speaking, and writing (including time for pretesting items, tutorials, and administrative questions). Working from there, we assume that the writing assessment portion of the TOEFL 2000 test could be allocated between 60 and 75 minutes. This time duration will include both the elaborated written response described in Assumption 1 above and brief written responses that may be

independent or interdependent with other test modalities such as reading and listening. This time expectation obviously constrains the extent and qualities of writing that can be sampled.

3. *The TOEFL 2000 writing assessment should elicit writing integral to academic environments and contexts.*

If the TOEFL test is to fulfill its purpose of informing admissions decisions in higher education, then the writing tasks should be situated in an academic environment. The challenge facing the TOEFL 2000 project is not to set entry-level standards beyond what is expected of native English-speaking students entering higher education. Rather, it is to provide a range of academically situated writing tasks to which examinees can respond and demonstrate their writing competence and from which those making admissions decisions can determine whether examinees are ready for full matriculation into the "academy" or are in need of additional writing and language instruction. In other words, we must distinguish between the language proficiency that an examinee has achieved prior to admission to an academic program and the language proficiency that is achieved through an enculturation in the "academy" (Carson, Chase, Gibson, & Hargrove, 1992; Ginther & Grant, 1996; Raimes, 1990; Waters, 1996).

The writing framework recognizes that writing is a communicative act and that communicative language ability must include contextualized language use (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980; Chapelle, Grabe, & Berns, 1997; Grabe & Kaplan, 1996; Hamp-Lyons & Kroll, 1997; Hymes, 1996; Savignon, 1983). This initial framework includes writing tasks that span a range of participants, content, settings, purposes, and registers outlined in the Jamieson et al. (1999) framework. In addition to more traditional essay writing tasks, we envision some short writing tasks such as e-mail exchanges or informal notes. However, we recognize that the academic writing that is valued most by both students and instructors is the writing that contributes directly to course grades (Anderson, Best, Black, Hurst, Miller, & Miller, 1990; Chiseri-Strater, 1991; Leki & Carson, 1994; McCarthy, 1987; Nelson, 1990). It remains for research in writing and the other skill domains to determine the extent to which various contextualized situations should be represented in the TOEFL 2000 writing assessment versus the other skill areas. Certain student-to-student exchanges, non-academic milieu, purposes (e.g., instrumental, regulatory, personal, interactional), and consultative and informal registers may be more appropriately reflected in the modalities of listening and speaking.

4. *The TOEFL 2000 writing assessment should select tasks thought to maximally generalize to and be recognized as important across disciplines.*

The research on writing in various academic disciplines cited earlier as well as recent reviews of needs analyses in English for Academic Purposes (EAP) research (Ginther & Grant, 1996; Waters, 1996) have identified and classified typical writing genres across a range of both undergraduate and graduate disciplines. Typical genres included summary writing, experimental (lab) reports, case studies, research papers, and book and article reviews. Several studies (Bridgeman & Carlson, 1983; Carson et al., 1992; Eblen, 1983; Hale et al., 1996) have also reported a high frequency of in-class writing tasks that included essay and short-answer exams where examinees were commonly expected to use the rhetorical forms of compare and contrast, analysis, and/or cause and effect. Ginther and Grant (1996) discuss the primary function of in-class writing as "providing an opportunity for teachers to evaluate and for students to display knowledge" (p. 11). Bereiter and Scardamalia (1987), Belcher (1995), and Cumming (1995)

further discuss academic writing that goes beyond displaying knowledge to critiquing, extending, and creating knowledge. However, Belcher notes that "a high level of domain expertise is needed to authoritatively and persuasively critique works in a specific discipline" (p. 136).

It would not be fair to assume that TOEFL examinees applying for undergraduate admissions have already been enculturated into a disciplinary discourse community, so it seems inappropriate for TOEFL 2000 writing tasks to impose cognitive demands that require such disciplinary knowledge or skills. It also seems inappropriate to include tasks that require primarily narrative or imaginative responses. Narrative and imaginative writing both tend to be associated primarily with English composition courses and have traditions that are highly culture-specific. Nonetheless, although the task stimuli we propose will not require examinees to generate primarily narrative or imaginative responses, they will leave open the option for examinees to utilize these rhetorical functions, if they so choose, in combination with other rhetorical functions in response to the TOEFL 2000 writing tasks.

5. *The TOEFL 2000 writing assessment should contain writing tasks that vary with respect to their dependence on reading and listening materials.*

A primary goal of the writing assessment is to move beyond a single essay as the measure of writing proficiency (Hamp-Lyons & Kroll, 1997). Additionally, the TOEFL 2000 writing framework recognizes that reading and writing are "inextricably linked" (Carson & Leki, 1993, p. 1) and that it is important to include writing tasks judged in terms of the writer's degree of content responsiveness. A considerable writer's body of research has examined writing in various academic disciplines and underscores the fact that academic writing rarely occurs as an isolated act but rather is often in response to the content of a course (Behrens, 1978; Braine, 1995; Bridgeman & Carlson, 1983; Carson et al., 1992; Eblen, 1983; Hale et al., 1996; Kroll, 1979; Leki & Carson, 1994; Ostler, 1980; Sherwood, 1977; Walvoord & McCarthy, 1990). However, evaluating writing integratively with other skill domains is problematic in that the assessment of writing is thereby confounded with other skills such as reading or listening. With integrated reading-to-write or listening-to-write tasks, successful performance on the writing task is dependent on successful completion of the reading and listening tasks. This is a critical issue when score users want information at the individual skill level (Grant & Ginther, 1995; Jamieson et al., 1999; Taylor, 1993). While Taylor, Eignor, Schedl, and DeVincenzi (1995) explain that interdependencies of this sort cannot be modeled with existing unidimensional or multidimensional item response theory (IRT) models used with the current TOEFL test, they note the work of Mislevy (1994a, 1994b) on an approach to modeling such interdependencies through the use of Bayesian Inference Networks. In this approach (Almond & Mislevy, 1997) a profile of an examinee's writing ability may be derived from a combination of independent writing tasks and a subset of interdependent tasks. Thus, with multiple writing tasks that include both independent and content-dependent tasks, we hope to move beyond the single, independent essay model to a writing model that is more reflective of writing in an academic environment while also dealing with the interdependency issue.

6. *The TOEFL 2000 writing assessment should place emphasis on (i.e., evaluate and score) written texts but also facilitate a variety of writing processes.*

Although evaluating examinees' composing processes is beyond the purview of the TOEFL 2000 test, the writing framework recognizes that writing is "multiply recursive rather than linear as a process"

Grabe & Kaplan, 1996, p. 87) and that examinees employ a range of writing processes in all text-making. The practical time constraints of the testing situation preclude many approaches to writing that highlight composing processes such as "invention and pre-writing tasks, and multiple drafting with feedback between drafts; a variety of feedback options from real audiences. . . , free writing and journal writing as alternative means of generating writing and developing written expression" (Grabe & Kaplan, 1996, p. 87). Grabe and Kaplan and others (Ginther & Grant, 1996; Hamp-Lyons, 1991; Hamp-Lyons & Kroll, 1997; Hughey, Wormuth, Hartfiel, & Jacobs, 1983; Johns, 1990; Krapels, 1990; Santos, 1992; Silva, 1990) have detailed the process approach movement in first and second language writing research and literature. Nonetheless, within the time constraints of the test, examinees should be given the opportunity to select and utilize writing processes appropriate to the task and testing situation. For example, some examinees may choose to mentally collect their thoughts and then write; others may select from a variety of recognized pre-writing activities such as mapping, scaffolding, listing, outlining, notetaking, or free writing. Some may choose to word process their responses while others may opt for handwritten responses. Examinees may also use part of their writing time to revise and edit their work. One approach we may consider as part of the research is to present the task(s) requiring an elaborated response early in the test and then allow examinees to return to their writing at the end of the testing period to review and edit their earlier work.

7. *The TOEFL 2000 writing assessment should identify writing tasks that will discriminate most at middle to upper levels of EFL/ESL writing proficiency.*

Weighing the time and cost constraints against the uses of test scores suggests that the writing tasks should target performances that will best inform decision-making about admissions to universities or colleges in North America. This means that very simple writing tasks such as copying or filling out forms will be excluded in order to include more complex, demanding writing tasks in response to which examinees can demonstrate their writing competence.

8. *The TOEFL 2000 writing assessment should score writing tasks based on characteristics of written texts that examinees produce as well as scorers' processes of reading those texts.*

In writing assessment, the evaluative criteria should be part and parcel of the questions and directives given to the examinee. Whether stated as part of the writing prompt proper or as preparatory material, there is a fundamental need to state and make clear that there is a reader/rater who is for some purpose receiving the written information and evaluating it in some way. A reader-writer model and aspects of the evaluation of language and discourse features are elaborated in the evaluative criteria section of this paper.

Organizing the Test Domain

Based on the objectives outlined in the previous section that identify the purpose and domain of the TOEFL 2000 writing assessment, we propose that this domain be organized according to three types of writing: (a) an independent invention task (i.e., a single essay based on a short prompt), (b) interdependent text-based tasks from reading and listening sources (e.g., a summary of a reading passage or listening lecture), and (c) interdependent situation-based tasks from listening or reading sources (e.g., an informal note to a professor). These types of writing are detailed in the following sections of the framework.

Identifying Task Characteristics

Thus far, we have identified the domain of the TOEFL 2000 writing assessment with respect to three types of writing. The next step is to decide which task characteristics to include in the writing framework. We have distinguished the characteristics of the writing tasks along three basic dimensions: (a) task stimuli designed to represent and elicit authentic examples of academic writing; (b) rhetorical functions reflective of informative writing; and (c) evaluative criteria based on a reader-writer model. The following sections elaborate these dimensions.

Identifying and Operationalizing the Variables

In operationalizing the task characteristics of task stimuli, rhetorical functions, and evaluative criteria, we further delimit the domain of writing, focusing specifically on the opportunities for examinees to demonstrate writing ability in an academic context and on the variables comprising readers' assessment of examinees' written texts.

Task Stimuli. In considering task stimuli, we realize that the types of prompts we propose here are necessarily provisional as we await research results. Nevertheless, we want to indicate our current thinking about the ways in which examinees might be asked to demonstrate proficiency in writing, understanding that the final set or combination of writing task types, arrived at through research and feasibility studies, might be different. At this point, the three types of prompts we have chosen to represent the task stimuli are: (a) an independent invention prompt designed to elicit a single essay; (b) an interdependent text-based prompt designed to elicit, for example, a summary of a reading passage; and (c) an interdependent situation-based prompt designed to elicit, for example, an informal note to a professor. Further details of these task stimuli are presented in the following subsections.

Independent Invention Essay

A task of independent invention will elicit an essay of, typically, four to six paragraphs. The prompt will provide just enough information to stimulate the examinee to generate his/her own ideas on the topic with supporting reasons and examples drawn from observation, experience, or reading. A task in this category will typically take the form of a two- to-four-sentence prompt requiring the examinee to do one of the following: (a) identify and explain a problem and then offer a solution; (b) take a position on an issue by agreeing or disagreeing with the point made in the prompt, or by selecting among alternatives provided in the prompt. Appendix B presents an example of a prototype of an independent invention task in which the examinee is asked to write a text presenting and supporting a position on the topic of whether technology makes life better or creates new problems. A prototypical essay response to the prompt is also provided.

Interdependent Text-based Response

An interdependent text-based task will elicit either a brief or elaborated written response. A brief response will consist of, typically, four or five sentences which will display the knowledge the examinee has derived from the text. A shorter task in this category will typically take the form of a reading passage

(although it could also be presented orally as an excerpt from a lecture), followed by reading (or listening) comprehension questions, followed in turn by a writing prompt requiring the examinee to identify, select, and manipulate the relevant information in the text. An elaborated response will likely consist of four to six paragraphs and require a critical consideration of the issue(s) addressed in the stimulus. Appendix C presents an example of a prototype of a text-based task (a reading passage on the control of fire by early humans), followed by a prototypical short-answer response to the prompt (the requirements for the successful use of fire by early humans).

Interdependent Situation-based Response

An interdependent situation-based task, which may be presented in oral or written form, will elicit a short response of, typically, four or five sentences appropriate to the content, setting, and purpose inherent in the prompt. A task in this category will typically take the form of an exchange between or among students or between a student and a faculty or staff member, which will then lead to a prompt requiring the examinee to do one of the following: (a) identify a problem, explain or elaborate on it, and either propose or request a solution to it; or (b) synthesize information and present, summarize, or report on it. Appendix D presents an example of a prototype of a situation-based task (a study-group discussion among three students), followed by a prototypical response to the prompt (a short note to the professor of the class).

Since an appropriate response to this prompt will be less formal than those of the previous two, register becomes a potentially significant variable. We recommend, therefore, that the question of informal expectations raised by this type of task be investigated as part of the research agenda.

To conclude this section on task stimuli, we reiterate the tentative nature of the types of prompts we have proposed and leave open to research such questions as, for example, whether a single sample of a longer piece of writing (e.g., an independent essay) would yield sufficient information about an examinee's writing ability, or whether a larger number of shorter samples (e.g., the interdependent tasks) would be more useful, or whether exclusively interdependent samples, longer and shorter, would work best. We regard these as issues to which research can inform final decisions about task stimuli.

Rhetorical Functions. The second dimension of the writing tasks for the TOEFL 2000 test is a core set of three rhetorical functions common to informative writing: categorization and analyses, problem-solution, and suasive argumentation. We have selected these rhetorical functions on the basis of their fundamental integrity to writing in North American university and college contexts, across a range of major academic domains. In doing so we have excluded rhetorical functions such as narration or creative expression that are less frequently required of students, are less integral to these situations, may be tangential to the core construct of academic writing, or which may produce biases or undue complications in administration of the test (cf. Hale et al., 1996; Hamp-Lyons & Kroll, 1997; Mohan, 1986; Waters, 1996). The rhetorical functions of each writing task in the TOEFL 2000 assessment might, for example, require examinees to either:

- categorize key features and/or analyze and describe relations between them;
- identify a problem and analyze it and/or propose a solution to it;
- or state a position, elaborate it, and/or justify it.

We consider these rhetorical functions to include related concepts such as definition, enumeration, or comparison/contrast, but the terminology above more generally applies across a range of writing tasks and the logical, cognitive, or functional processes integral to them. Importantly to the test context, we envision these core rhetorical functions as forming a basis for designing writing tasks that sample examinees' writing across each of these rhetorical functions, that span and interrelate particular task stimuli (i.e., independent invention essays, brief situation-based responses, and summary writing) in simple or more elaborated or complex forms. This should allow for a maximum amount of information to be obtained about each examinee's writing abilities while sampling across a variety of types and durations of writing. Moreover, at the level of rhetorical function, this information can be compared or verified for each examinee across the writing tasks that he or she performs.

Topic Characteristics. Topic characteristics need to be considered as well in the design, verification, and validation of the writing tasks for the TOEFL 2000 test. These will be especially important in ensuring the comparability or equating of specific writing tasks across different versions of the test, and of course in terms of the accessibility or fairness of the test. Although this issue has received little research, most of the studies that have tried to investigate topic variables in writing have found them to be a significant source of variation in students' writing performance or the design of writing tasks (cf. Brossell, 1986; Hidi & McLaren, 1990; Hoetker, 1982; Ruth & Murphy, 1988; Smith et al., 1985; Tobias, 1994). Important characteristics of topics related to writing tasks that may also characterize aspects of task difficulty include:

- academic contexts and content,
- the extent of required topic-specific background knowledge and personal experience of examinees,
- the emotional appeal or interestingness of a topic to examinees,
- the extent of options or choices available to examinees in their writing,
- the extent of topic abstractness or specificity,
- the differences in examinees' cultural or geographical backgrounds,
- the cognitive demands of a topic,
- the information load of the topic (i.e., the degree of task structure provided and the extent to which the task specifies a required role, audience, purpose, form, or mode of discourse),
- time elements,
- the length of response required, and
- access to source material(s).

Evaluative Criteria. Meaning can be conveyed in various ways in written texts; conversely, particular deficiencies may hinder its conveyance. Many specific variables, or components of them, influence the effectiveness of written communication, each of them interdependent with the others in fundamental but subtle ways. Organizing and presenting ideas in a written text depends, for example, on the selection of appropriate words and phrases; on facility with the conventions of grammar, punctuation, and spelling; and on the competent use of logic and rhetorical devices to sustain a reader's attention and direction. The realizations of these variables, in turn, vary according to context, purpose, and situation. The sources of difficulty that they imply are mostly speculative at this point insofar as any generalities might be made about ESL writing overall, or even for particular types of writing tasks.

The nature of these complex variables can be viewed through two complementary orientations that commonly appear in evaluation practices, research, and theories concerning adults' writing in second languages in academic contexts. One orientation considers the characteristics of the written texts that people produce: What text features characterize examinees' writing in a second language? The second approach considers the perceptions and judgments of readers of such texts: How do readers of second-language writing think and respond while they read examinees' writing? We can call these two orientations a text-characteristics model and a reader-writer model. Each is useful in conceptualizing how writing tasks for the TOEFL 2000 test might be scored and scaled.

Text Characteristics Model

Although considerable research on the characteristics of ESL writing has emerged over the past two decades, this research has mostly addressed very specific aspects of written texts or issues local to single contexts (Cumming, in press; Cumming, 1998; Silva, 1993). Little inquiry has tried to take a systematic, comprehensive perspective on all of the variables that combine generally to make ESL texts more or less effective. Reviewing these studies, collectively, points toward some of the evaluative criteria that might form a basis for scoring and scaling TOEFL 2000 writing tasks. These studies have highlighted either a macro-level perspective on ESL written texts, addressing adult ESL students' organization of written discourse and presentation of ideas, or a micro-level perspective, addressing aspects of the syntax, morphology, or lexis that appear in adult ESL students' writing.

Discourse and ideas. Among the text characteristics identified in published empirical studies considering the organization of discourse and ideas in compositions written by adult ESL learners are variables such as:

- hierarchy of related ideas that specify relationships between major ideas and less important ones, links of superordination and subordination, stated facts and supporting information to form a visible macro-structure and micro-structure throughout the text developed to achieve relevant, appropriate pragmatic effects (Chan, 1997; Whalen & Menard, 1995; see also Freedman & Pringle, 1980);

- introductory framing, such as thesis or viewpoint statements in argumentative writing, in the initial part of the text that is sufficiently visible, signaled, and coherent to allow readers to make appropriate predictions about the sequence and organization of ideas in the text (Tedick & Mathison, 1995) or to indicate that the writer is trying to convince a reader of specific ideas

(Kobayashi & Rinnert, 1996), and concluding phrases to the text that provide the reader with a definite sense of closure and completion (Chan, 1997); and

- appropriate, accurate, and extensive paragraphing and other coherence breaks (subdivisions of text chunks) as well as rephrasing, pronouns, synonyms, repetition of phrases and other inter-sentential connectives, links, or semantic chains sufficient to maintain thematic coherence, develop the topic cohesively, and orient readers, without frequent digressions, redundancy, or disruptions to the flow of reading (Chan, 1997; Intaraprawat & Steffensen, 1995; Reid, 1992; Reynolds, 1995; Wikborg, 1990).

Language use. Published empirical studies that have addressed matters of language use in adult ESL students' written texts have highlighted such variables as:

- the total number, range, or overall quality of words written in relation to the time available, the task, or threshold levels of vocabulary established in relation to word frequency counts of corpora of English texts or to readers' impressions of the vocabulary (Cumming & Mellow, 1996; Engber, 1995; Frase, Faletti, Ginther, & Grant,1997; Kroll, 1990; Laufer & Nation, 1995; Sasaki & Hirose, 1996; Santos, 1988; Sweedler-Brown, 1993; see also Connors & Lunsford, 1988; Vann, Meyer, & Lorenz, 1984);

- appropriate, accurate, extensive use of key lexical items with important rhetorical functions, such as illocutionary markers to indicate the illocutionary act that the writer is performing, doing so in a manner that makes assertions that are warranted, appropriate, and not overly obvious (e.g., "I suggest that," "In sum," "My hypothesis is," "I claim that," etc.; Allison, 1995; Intaraprawat & Steffensen, 1995), or indicators of epistemic certainty, probability, possibility, and usuality and other kinds of hedges (i.e., uses of words such as "certainly," "indeed," "in fact," "believe," "seem," "appear," "would," "may," "might," "perhaps," "possibly," "often," "usually," etc.; but not overuse of words such as "actually," "in fact," "know," "think," "will," "may," "always," "usually," etc.; Hyland & Milton, 1997; Intaraprawat & Steffensen, 1995; see also Zhang, 1995);

- appropriate selection, use of, and distinctions between paraphrasing, restating, citing, and quoting source material, as well as evaluating or querying it appropriately and critically, as well as indicating, explaining, or stating one's personal perspective about it (i.e., in writing tasks that rely extensively on reading material; Braine, 1995; Campbell, 1990; Carson et al., 1997; Connor & Kramer, 1995; Cumming, Rebuffot, & Ledwell, 1989; Deckert, 1993; Dong, 1996; Johns, 1985; Johns & Mayes, 1990; Sarig, 1993);

- appropriate, accurate, extensive use of subordinate clauses and participle phrases of considerable length, range, and diversity (Dickson, Boyce, Lee, Portal, Smith, & Kendall, 1987; Homburg, 1984; Ishikawa, 1995; Perkins, 1980; Polio, 1997; Reid, 1992);

- appropriate, accurate, extensive use of specific grammatical features such as articles (i.e., a, an, the, and ∅ article) in noun phrases (Bardovi-Harlig & Bofman, 1989; Cumming & Mellow, 1996; Janopoulos, 1992; Kroll, 1990; Master, 1986; Polio, 1997; Sweedler-Brown, 1993) or the present perfect with appropriate temporal adverbials (e.g., at present, up til now, so far, lately, since-phrases, recently, already, yet, etc.; Bardovi-Harlig, 1997; Janopoulos, 1992; Sweedler-Brown, 1993);

- appropriate, accurate punctuation (particularly commas to signal introductory elements, relative and non-relative clauses, series, participle phrases, coordination, and subordination; capitalization of proper nouns; quotation marks; semi-colons; possessive apostrophes; Connors & Lunsford, 1988; Janopoulos, 1992; Kroll, 1990; Sweedler-Brown, 1993); and

- appropriate, accurate spelling (to a certain threshold or frequency level; Dickson et al., 1987; Frase et al., 1997; Janopoulos, 1992; Sweedler-Brown, 1993).

Merely suggestive at this stage, these text characteristics point toward specific kinds of evaluative criteria, together with additional variables, that might be assessed more rigorously in developing tasks for the writing component of the TOEFL 2000 test. Although each text characteristic has some empirical basis and theoretical justification (in the publications cited), none can be considered thoroughly validated or uniquely integral to the full range of ESL writing at this time. Moreover, these characteristics have been derived from specific sets of texts written by particular populations in particular contexts, so their potential to generalize to other texts, populations, or contexts is probably quite limited.

Reader-Writer Model

From this perspective, evaluative criteria can be stipulated for ESL writing from the viewpoint of the people reading such texts. Research from this perspective has adopted several orientations. One strand of studies has surveyed relevant readers of ESL written compositions, such as professors or ESL instructors, to identify the aspects of ESL written texts they consider most effective or important, and to determine how they may vary. Such studies have contrasted the evaluation judgments of particular groups, such as ESL instructors, professors in particular disciplines, or raters with distinct cultural backgrounds (e.g., Brown, 1991; Kobayashi & Rinnert, 1996; Mendelsohn & Cumming, 1987). Related studies have surveyed professors to identify the errors in ESL students' writing they find the most irritating, calling this "error gravity" (e.g., Rifkin & Roberts, 1995; Santos, 1988; Vann, Meyer, & Lorenz, 1984). Focused on evaluation practices, numerous studies have, following trends reviewed by Connor-Linton (1995) and Huot (1990), inquired into the thinking and decision-making processes that raters of ESL compositions use to guide their assessments. Generally, such inquiry has found raters using complex forms of decision-making, reading ESL texts to sample for key text characteristics and issues of interpretability, weighing the value and interactions among these to reach a scoring decision. Although descriptions of the mental processes used to rate ESL compositions have been documented in such inquiry, it is also apparent that the nature of the evaluative criteria that raters use, and the assessment decisions they make, depend on such factors as training they may have received or a particular rating scale they use (Cushing-Weigle, 1994), their experience or skill doing such assessment (Cumming, 1990), and cultural values or expectations of their academic disciplines (cf., surveys cited above of different rater populations). For this reason, it is not possible to specify a universal model of reading ESL compositions, though this is a primary area for research and development in the context of developing and validating writing tasks and scales for the TOEFL 2000 test.

Moreover, such inquiry makes clear that evaluative criteria need to be stipulated precisely in the instructions and directives given to examinees taking the TOEFL 2000 test. Whether stated as part of a prompt for writing tasks or as preparatory material orienting examinees, there is a fundamental need to make clear to examinees that they are writing in the TOEFL 2000 assessment for a reader who will read, interpret, and evaluate their written texts with specific purposes and conceptualizations. For example, for

the traditional "essay," or what we call the Independent Invention task, directives and materials about the test should be given to examinees so that they can assume that readers of their writing:

1. know the prompt and task they have been assigned;

2. know how much time has been allotted to the task and have some expectations about the amount and specificity of development that can be expected of writers at various levels;

3. are interested to know the writers' perspective (e.g., from the perspective of an "educated non-specialist audience");

4. have no particular biases on the topic or issue (unless the examinee through the directives is instructed to anticipate a reader with a particular stance);

5. are proficient readers of English, who will read a text in one pass, not expecting to have to retrace any steps or be slowed down unduly while reading;

6. would notice information that does not seem to be advancing the exposition on the topic or that would be inconsistent or redundant;

7. will make effort to extract meaning from text that is not fully grammatical or correctly punctuated, though such effort will not be heroic;

8. will note where failings of syntax, word choice, and other aspects of language use do or do not affect meaning in a major way; and

9. will observe and note almost all violations of spelling and conventions of standard English but will also be aware of regional differences.

In heavily content-dependent tasks, such as summaries or even tasks that ask for literal conveyance of essential information in a chart, graph, or schematic, the instructions in the exam should lead the examinee to assume that readers of the written text:

1. have fairly precise expectations about the level of specificity of the content expected to be covered as well as the sufficiency of information if not the essential propositions expected to be reported or covered;

2. know the source material but cannot be expected to draw on this knowledge to fill in anything left out by the writer; and

3. have expectations about whatever directives are given about paraphrase, citations, quoting, and other means of indicating sources and distinguishing such source citation from personal opinion if that is appropriate to the task.

A Preliminary Synthesis of These Evaluative Criteria. Given the limited state of knowledge about task characteristics and readers' processes of rating ESL compositions, any specification of evaluative criteria for writing tasks for the TOEFL test is necessarily tentative. Indeed, specifying, refining, and

validating such criteria has to be a major research and development undertaking done in the preliminary phases of work on the TOEFL 2000 assessment. For the purposes of suggesting how such work might proceed, we can outline a preliminary set of operational criteria to be investigated, using the independent invention essay as a prototypical example. Evaluative criteria we believe are best considered by raters analytically scoring two dimensions of examinees' writing: (a) organization of discourse, ideas, and other substantive content to fulfill a specific writing task and (b) accurate and appropriate use of the English language and conventions of written English. A major research project, however, is needed initially to determine whether there is sufficient difference in TOEFL examinees' performance on these two dimensions to justify their being scored separately. The TOEFL 2000 speaking framework (Butler, Eignor, Jones, McNamara, & Suomi, 1999, pp. 13-15) outlines a set of important considerations and options to be considered in devising appropriate scoring methods and criteria for speaking tasks. These suggestions apply equally to the TOEFL 2000 writing tasks. We envision a multi-trait scoring procedure for the writing tasks (including at least the two traits of discourse and ideas and of language use). This could function as an overarching scheme for scoring, supplemented by specific variables particularly salient in specific writing tasks (as may be illuminated and verified by our research) as well as by guidelines for binary choices among these variables to guide and focus raters' decisions during scoring. However, considerable research is needed to specify these empirically as well as to verify their suitability and validity.

Organization of Discourse and Ideas for Independent Invention Essays

This aspect of writing includes components of organization, coherence, and progression, and development of the discourse and its ideas.

Organization: The reader is consistently able (i.e., on-line, without undue rereading) to follow and understand what the writer is discussing and why it is being discussed in relation to the prompt given or in relation to other considerations that the writer has already felicitously communicated to the reader and are relevant to the prompt.

Coherence: The reader can derive meaning from sentences and not find logical contradictions in or among them; the reader does not encounter information that seems irrelevant; the reader is not required to fill in or guess at unstated facts, supports, presuppositions, or premises beyond those that would ordinarily and more or less automatically be filled in by a native reader of English reading competently written standard English prose.

Progression: The reader does not encounter glaring gaps, redundancies, or digressions in the flow of the text.

Development, specificity, and quality of information and ideas: The writer is judged to have presented fully specified and relevant details, examples, reasons, and reasoning that bear on and respond to the task and topic; these go beyond merely listing, enumerating, or abstractly alluding to points so as to specify fully, clearly, and convincingly the information conveyed. Depending on expectations indicated in the scoring rubric and prompt, as well as cultural norms and expectations, the quality of ideas conveyed may be thoughtful, compelling, convincing, or unique.

In addition to these criteria, for writing tasks that are extensively dependent on the content of reading or listening material, the accuracy of information and ideas, as well as sufficiency of coverage of them,

needs to be determined and specified, for example, in terms of an expected range of information given the prompt. For tasks that ask for summarization or description, evaluation needs to consider how much of this expected content is included in the written composition, how it is organized and conveyed, and what degrees of relevancy, specificity, and accuracy are expected, as well as expectations for paraphrase, citation, or quotation.

Language Use

Following the list of variables itemized on pages 15 and 16, evaluation of language use might consider specific linguistic variables salient in examinees' written texts, such as can be verified through research on the TOEFL 2000 tasks. Some may be amenable to computer-generated analyses (e.g., word counts, word frequency counts in relation to threshold levels of vocabulary knowledge, and perhaps analyses of syntactic structures) to inform raters' judgments. Each variable, however, will be relative to the task type, topic variables, and overall fluency of the piece:

- appropriacy and range of vocabulary and idiom used;
- appropriacy and effectiveness of language used to manage discourse connections;
- accuracy, appropriacy, and range of phrase-level syntax and morphological markers; and
- spelling, punctuation, and other orthographic or typographic conventions, and for tasks directly dependent on content from reading or listening passages, a measure of ability to use specific language elements in paraphrase, reported speech, and summation.

Appendix E presents the TOEFL 2000 writing variables categorized according to Almond and Mislevy (1997): (a) variables that delineate the domain, define features for constructing tasks, control for test assembly, and characterize aspects of task difficulty and (b) those that characterize performance or individual responses (i.e., scoring rubrics and rules for scoring).

4. Research Agenda

An overarching goal of our research framework is to establish the validity of the interpretations, and the appropriateness of the actions, that result from a TOEFL 2000 writing test. Two objectives under this goal are to:

1. Examine the extent to which the test as a whole and each of its constituent tasks adequately represent the conception of writing that we have set forth. That is, which aspects of the multi-faceted conception are best and least well captured by the measure, and what are the consequences of any underrepresentation? What processes, strategies, and knowledge (as set forth in our conception of writing proficiency) are actually implicated in the performances elicited by TOEFL 2000 writing tasks?

2. Determine the relative influence on task difficulty and test performance of construct-irrelevant factors. What are the consequences of these extraneous factors in terms of both the interpretation of test scores and the consequence of their use?

We believe that these objectives can be achieved by mounting research and development activities under several broad categories: (a) refining the construct, (b) establishing sources of *task* difficulty, (c) establishing *score* meaning, (d) determining score utilization and score utility, and (e) other ancillary activities. Research activities can be distinguished according to (a) when they can be expected to occur, (b) whether they are more developmental/formative or summative in nature, and (c) the size or scale of the effort involved. We envision that small-scale prototyping and piloting will be carried out first, followed by experimental pretesting and, finally, large-scale data collection in operational settings — either in conjunction with or as part of actual operational administrations. Furthermore, we anticipate that several activities will be repeated over the course of development of the measure, first in small-scale, exploratory efforts and eventually in larger-scale, more formal studies.

Refining the Construct

A Reality Check. Conferring early on with a variety of constituencies (e.g., through small, select focus groups) should prove useful in several ways. First, gathering preliminary reactions of potential test users and language experts to prototype tasks will help to establish the credibility/acceptability of the tasks themselves. Allowing constituents to react to actual tasks should result in a greater sense of how our conception of writing corresponds to the notions held by those who are likely to have a stake in the TOEFL 2000 writing assessment. These early reactions will help to guide further development and research. Perceptions of the likely washback effect of various item types and formats can be obtained at this early stage also. Here, the reactions of potential test takers may be useful also. Whichever (and whenever) constituencies are consulted, it will be important that we have some faith in the prototypes that we put forth, lest we risk undermining constituents' confidence in our endeavor. The development process will undoubtedly be iterative, entailing several rounds of review and revision, and it will be important to carefully plan the timing of these cycles.

Guiding Research. Reactions of potential test users (and possibly test takers) can also be helpful in identifying possibly important construct-irrelevant sources of variance in test scores as well as ways in which important facets of writing proficiency may not be adequately represented in our prototype tasks. This information can be used to establish priorities for studies of the consequences of construct underrepresentation and of construct-irrelevant factors. As an example, if we encounter widespread

concern that our tasks do not permit writers to demonstrate all of the important aspects of the writing process (planning or reflection, for example) or that they may be disadvantageous to certain kinds of writers, then research can be designed to assess the effects of these perceived shortcomings.

Establishing Sources of Task Difficulty

We envision that the analysis of task difficulty for writing (and speaking) will require somewhat different, and perhaps untried, techniques than those that have been used so successfully to model task difficulty for reading. Although the features of essay prompts and topics may be amenable to such analyses, the stimuli (or prompts) for writing tasks will undoubtedly be much more limited than those (reading passages) employed to assess reading skills. The main target of analysis for writing, therefore, will be the text that examinees *produce*, rather than the text to which they *respond*.

Literature Review. Constituents' reactions may provide some insight into variables that underlie task difficulty. A review of relevant literature (already underway in the context of the present report) will be even more useful, however. Numerous linguistic variables have been identified in the research literature as hallmarks of good writing or as factors that either hinder or facilitate the transmittal of meaning in written text. As suggested earlier, the influence of many of the variables that have been mentioned is still quite speculative, and there have been relatively few studies that have investigated the joint influence of combinations of variables or the possible interactions among them. Results of reviews for the other three language modalities should also prove useful, as many of the factors that determine difficulty for reading, for example, will also be the same ones that relate to the quality of writing.

Studying Alternative Tasks. Two critical variables for the writing component of the TOEFL 2000 test will be the *length* of the responses that are elicited from test takers and the degree to which they depend on stimuli or require independent invention. Clearly, a tradeoff exists between the number of tasks that can be administered and the length of the responses that can be obtained. One important question, therefore, concerns the appropriate mix of extended versus shorter responses (and, indeed, what in fact constitutes an extended response in the minds of constituents and writing experts). The degree to which several short responses can capture the same qualities as longer pieces, for both independent invention and interdependent text-based tasks, is an empirical question.

Toward Models of the Reader and the Text Characteristics. A series of studies would prove useful to determine the extent to which each of a wide variety of linguistic variables affects raters' judgments of performance on alternative writing tasks. The question of interest here is "What features do raters attend to, and what weight do they give to them?" This determination could be made in the context of a series of experimental studies that systematically manipulate, in various combinations, the features of writers' responses. The objective of these "policy capturing" studies would be to ascertain the variables that most highly correlate with, and presumably influence, readers' judgments of the quality of examinees' writing. Analyses could be carried out for different kinds of readers who vary with respect to English language facility, academic background, motivation to understand the writer, etc. in order to work toward a reader model of writing. Specific audiences could represent faculty members, fellow students, ESL experts, and so on. Convergence by multiple audiences on the same linguistic variables would have implications for developing scoring guides and, perhaps, for anchoring the meaning of test performances. In addition, a model of texts could be developed by analyzing readers' reactions to and interpretation/understanding of texts having varying qualities of language.

In addition to including the consideration of the judgments of alternative kinds of readers or audiences, it will be informative to obtain a variety of different reader reactions such as:

- ratings of various kinds, for example

  -- degree to which a task was fulfilled by the writer

  -- amount of effort required to read and understand a writer

  -- estimates of a writer's chances of "academic survival"

- reader behaviors

  -- time required to read

  -- need to reread

Alternative Scoring Methods. Another approach to establishing the difficulty of writing tasks would be to investigate the application of alternative scoring methods/rubrics/standards in order to determine their effect on difficulty, reliability, and validity. For subjective judgments (such as the kinds of ratings typically used in evaluations of writing), difficulty may depend considerably on the nature of the judgments that are made. The particular features that are specified in scoring schemes and the stringency with which scoring criteria are applied may, in a very real sense, be important determinants of performance expectations. Decisions need to be made regarding the optimal approach to scoring (i.e., holistic, multi-trait, or primary trait methods), the precise criteria to be scored (and their construct relevance), and the procedures that raters will follow to do the scoring (e.g., binary choices, individual or group scoring sessions, calibration to model compositions). Research is needed with sample tasks to determine what is most reliable, informative, valid, and feasible concerning these matters.

Natural Language Processing. Finally, we suggest the need for a vigorous program of research on the application of natural language processing (NLP) methods to the scoring of writing tasks. Currently, methods are being developed at ETS to identify a variety of rhetorical, syntactical, and lexical features of essays (Burstein et al., forthcoming; Burstein, Kukich, Wolff, & Lu, 1997). These features, in combination, have been shown in preliminary research to be relatively strong predictors of the ratings given by human readers. These techniques should be extended and applied to the prototype tasks developed for the TOEFL 2000 writing measure. NLP methods, when applied to texts produced by examinees, should provide a good complement to policy-capturing studies, in which examinees' texts are systematically modified or manipulated. Additional reflections on the utility of NLP methods, as well as on hardware and software issues related to accommodating writing processes and capturing handwriting, are contained in Appendix F.

Establishing Score Meaning

Construct-irrelevant Influences. When the most likely construct-irrelevant sources of test-score variance have been identified and prioritized, studies will be needed to assess the plausibility (and magnitude) of the threats that these factors represent. As one example, study will be required of the

extent to which differential facility with word-processing (and with various aspects of the computer interface) may affect test performances. Speededness is another perennial threat to test score validity that should be studied. We will need, therefore, to determine the consequences of imposing strict time limits on various groups of examinees. Also, because our conception of writing emphasizes writing skills over content and background knowledge, it will be desirable to ascertain the degree to which examinee variability in the familiarity with task topics may influence test performance. These are just a few of the possible sources of extraneous test variance that should be considered.

Internal Test Structure. The internal structure of the TOEFL 2000 test battery should be studied to ascertain the relationships among the various test sections and tasks (reading, listening, speaking, and writing). Questions to be addressed here include:

- Does a minimum threshold level of performance on some skills (e.g., reading) seem necessary for adequate performance on others (e.g., writing)?
- Do the writing tasks relate to one another in theoretically expected ways? For example, do tasks that require varying degrees of independent invention relate to one another in anticipated ways?
- Do performances on individual writing tasks relate more strongly to one another than to performances on other, non-writing tasks?
- Do integrated tasks relate in expected ways to independent tasks?

External Relationships. The meaning of TOEFL 2000 writing scores should also be established in relation to other external variables and measures. These variables might include the current TOEFL test (and its constituent parts), the Test of Written English, and other prominent language tests. Because, however, the TOEFL 2000 assessment measures are intended to be better than the current TOEFL test, the relationship should be less than perfect. Other important, well-established "marker" variables of specific language skills (e.g., appropriate and accurate use of devices to maintain thematic coherence) and relevant cognitive abilities (e.g., verbal fluency) should also be considered. Some of these latter markers might need to be developed. Other external measures might include admissions measures such as the new GRE writing test and the SAT® II writing assessment.

*Group Comparisons***.** Comparisons among various, relevant groups of test takers will be a useful way to establish score meaning. Among those comparisons that may prove useful are the following:

- graduate vs. undergraduate students
- native vs. nonnative writers
- skilled vs. less skilled writers (in their native language)
- before and after specific relevant instructions
- over time as reflective of more general experiences, including having taken the test previously

In addition to examination of the test performances of different groups, think-aloud studies of writers as they perform different tasks may prove useful in determining the relevant and irrelevant processes that examinees use when they take the test.

Generalizability. Typically, tests that require examinees to produce extended responses (including those designed to assess writing skills) face a significant tradeoff: usually only a limited number of tasks can be posed in the amount of time available for testing. Often, examinee performance does not generalize over alternative tasks designed to measure a given aspect of the construct of interest. That is, an individual's performance may depend to a considerable extent on the particular tasks that he or she is asked to do. Estimates will be needed, therefore, of the generalizability of scores derived from various combinations of writing tasks in order to shape decisions about the most appropriate test configuration. To the extent possible, it will be extremely useful also to try to understand the features of individual task types that contribute to any lack of generalizability that is detected. Formal study will be needed, therefore, to determine how much variation can be expected with respect to examinees, tasks, and person-by-task interaction for each of several different configurations of a TOEFL 2000 writing measure.

Determining Test Score Utilization and Utility

Consequences of Test Introduction. Once the test is operational, it will be highly desirable to ascertain how test scores are being used and to learn about both desirable and unintended effects of their use on institutions and on test takers. Some of these effects might be anticipated early in the development process by consulting potential test users and test takers. Eventually, though, formal surveys should be undertaken to determine (a) how test scores are employed in decision-making and (b) how examinees go about preparing to take the test.

Alternative Uses. Study will also be needed to assess the utility of test scores for different uses:

- for assessing ability to use English in an academic setting in order to fully participate in academic programs (admissions decisions),
- for decisions about course placement,
- for decisions about instruction, and
- for gauging progress, either individual or group.

Each of these uses, as well as any others that may arise, will require different kinds of evidence and different study designs.

Other Innovations

In conjunction with research focused specifically on the design and development of the TOEFL 2000 writing measure, it may be desirable to conduct more fundamental research on some of the generic task types (e.g., summarization, paraphrasing) upon which the TOEFL 2000 tasks may be based. This research would position the program to make further, state-of-the-art improvements in the future.

Some other possible innovations in testing might also be the subject of research. For example, there is considerable interest in some quarters in allowing examinees to choose the test tasks they undertake. Permitting such choice has a number of putative advantages, particularly when a limited number of tasks can be undertaken. Existing research has shown, however, that because, for instance, examinees do not always make wise choices, fairness is an important issue to consider when permitting examinee choice. Research could provide information about the circumstances under which choice might improve measurement, as well as the conditions under which it would be inappropriate.

The prepublication of pools of test items, particularly essay prompts, is another example of a testing procedure that has both potential advantages and disadvantages. Although one school of thought suggests that writing scores will exhibit greater validity if examinees have had an opportunity to reflect on the prompts they may encounter on the exam, others feel that predisclosure may invalidate test results insofar as some test takers may be inclined to prepare "canned" responses and simply reproduce them on the exam. This too is a researchable issue.

5. A Better Writing Test

What will make the TOEFL 2000 writing component a better writing test?

The writing assessment planned for the TOEFL 2000 test represents a significant improvement over the current TWE test, building productively on more than a decade of experience with the TWE test and related tests of writing at ETS (e.g., GRE, GMAT, NAEP) as well as considerable recent inquiry into writing practices and assessments in academic contexts. The research conducted as part of the TOEFL 2000 project will provide evidence of the extent to which our anticipated improvements have been realized.

We anticipate the following principal advances:

1. multiple writing tasks;

2. integration of writing tasks with related communication skills (e.g., reading, listening, speaking) and rhetorical functions (e.g., summarizing, persuading, describing);

3. more precise definition and scaling of the construct of writing in English as a second/foreign language for academic purposes, following a systematic program of validation research;

4. more extensive information on examinees' performance which better informs decision-making by score users; and

5. improved washback in English as a second/foreign language instructional programs.

These features will provide the test with enhanced educational relevance, more reliable assessment methods, validation evidence to inform the value and utility of the test as well as interpretations of its results, and a more positive impact on English language instruction. Moreover, they will provide a greater degree of explicitness about the assessment of examinees' writing not presently possible with the current TWE or TOEFL tests. They will help test users and examinees know more precisely what the writing component of the TOEFL 2000 test does and why. They will describe with increased precision the qualities of writing in English that examinees produce for a range of text types integral to academic performance.

Multiple Writing Tasks

Many current tests of ESL/EFL writing (including the TWE test) require examinees to produce a single written composition. This is problematic because only a single score for writing is elicited from each examinee, and only a single sample of writing is taken to be representative of the diverse domain of academic writing. In contrast, for the TOEFL 2000 writing test, examinees will produce a range of written texts, including an extended essay plus shorter responses. Multiple writing tasks offer several obvious advantages over existing ESL/EFL writing tests. In terms of sampling examinees' performance, an increased range of types of writing will be elicited, allowing examinees to demonstrate their writing abilities in various relevant ways. In terms of scoring reliability, results from the various writing tasks can be compared or aggregated, providing multiple sources of information to discern and confirm the range of examinees' abilities. In terms of construct relevance, examinees' writing will be assessed across a range of diverse task types that involve different ways of organizing and conveying information and

different situational factors, both integral to writing in academic contexts. This will improve the relevance of the test for educators and for educational purposes.

Integration of Writing with Other Academic Skills

The TOEFL 2000 test will treat writing as both an independent activity (requiring examinees to generate an essay that presents their own ideas) and an interdependent activity (requiring examinees to respond appropriately in writing to text-based and situation-based prompts that are dependent on reading or listening stimuli). The main advantage of doing this is increased authenticity. Writing in academic contexts is both an independent and interdependent, situationally embedded activity. Its effectiveness is judged as such by professors grading papers and other written products. As well, students produce written texts to convey information independently and in response to the topical content of information presented in other modalities (such as reading and listening). By organizing writing assessment from this perspective, the TOEFL 2000 writing assessment will evaluate examinees' abilities realistically with respect to the demands of academic writing and appropriately with respect to examinees' handling of information in other modes of communication. The TOEFL 2000 assessment will further distinguish the writing performances by separately scoring the responses to independent and interdependent writing tasks. Moreover, the TOEFL 2000 writing component will assess examinees' writing performance on a range of rhetorical functions integral to academic contexts (e.g., summarizing, reporting, categorizing, analyzing). The TOEFL 2000 writing framework thus expands the construct of writing to be assessed, aligning it appropriately with the kinds of writing that are integral to academic settings and evaluative criteria used to judge the effectiveness of the written products.

Definition and Scaling of the Writing Construct

Perhaps the greatest step forward in the TOEFL 2000 writing assessment is the potential to reach precise definitions of the range of abilities that form the construct of ESL/EFL writing. Although educators around the world regularly work with implicit understandings of what constitutes effective English writing, no existing research or testing programs have proposed or verified a specific model of this, such as would be universally accepted. Indeed, current ESL/EFL writing tests operate with generic rating scales that can reliably guide the scoring of compositions but which fail to define the exact attributes of examinees' texts or the precise basis on which they vary from one another. The opportunity to arrive at a general definition of the ESL/EFL writing construct in an academic setting, and to validate it empirically, appears feasible in the context of the TOEFL 2000 project, given the large pool of examinees internationally, the planned scoring procedures, and the proposed research program.

Scaling the range of examinees' performance on a core set of writing tasks in the TOEFL 2000 test will permit (indeed require for the purposes of test validation) two fundamental questions to be addressed: (a) Which qualities of examinee-produced written texts generally distinguish their effectiveness? (b) Which qualities of these texts do raters attribute chief importance to in judging their effectiveness? Research to answer these fundamental questions will produce precise models, scaled for ranges of task difficulty, ESL/EFL text qualities, and rating processes. Moreover, these models will be substantiated in respect to key sources of variation, such as the relative independence or interdependence of writing with other modalities of communication and specific rhetorical functions integral to academic performance. The outcome of the TOEFL 2000 assessment validation will assure that the test is

measuring what it is intended to measure. It will also establish normative standards for assessing the qualities of ESL/EFL writing, a step of enormous value for informing educational policy and practice.

Information about Examinees' Writing Abilities

The TOEFL 2000 test will produce a single score for each examinee's writing performance, representing the person's position on a normative scale related to all other TOEFL examinees intending to study at universities or colleges in North America. From this, score users will more clearly and precisely see where an examinee's writing skills stand in relation to those of other examinees. Above and beyond this, however, the TOEFL 2000 writing test will have the potential of providing score users and examinees with additional information about each examinee's writing. This information may be useful for instruction, learning, or diagnostic purposes. One set of information will allow a comparison of an examinee's strengths and limitations when writing on three types of tasks: (a) an independent invention prompt designed to elicit a single essay, (b) an interdependent text-based prompt designed to elicit, for example, a summary of an academic discussion, and (c) an interdependent situation-based prompt designed to elicit, for example, an informal e-mail message to a classmate. A second set of information will allow a distinction of an examinee's performance in writing versus other communication modalities, specifically reading, listening, and speaking. A third set of information will describe how well an examinee is able to use writing to perform rhetorical functions integral to academic tasks, such as reporting, categorizing, or synthesizing.

Washback

The washback effect, sometimes referred to as the systemic or consequential validity of a test (Alderson & Wall, 1993; Bailey, in press; Messick, 1989, 1994), of the current TWE and TOEFL tests has generally been viewed as negative (Bailey, in press; Hamp-Lyons & Kroll, 1997; Pierce, 1992; Raimes, 1990). The TOEFL 2000 writing test will go well beyond the single essay required by the TWE test by providing multiple writing tasks, both independent and interdependent tasks, more precise definitions and scaling of the writing construct, and more extensive information about examinee performances. Thus, the TOEFL 2000 writing assessment will provide the opportunity for the test to have a more positive impact on ESL/EFL pedagogy and curriculum development. Messick (1996) states that language tests which promote positive washback are likely to include tasks which are "authentic and direct samples of the communicative behaviors of listening, speaking, reading, and writing of the language being learnt" (p. 241). The research carried out as part of the TOEFL 2000 project will provide empirical evidence of the extent to which the TOEFL 2000 writing tasks reflect desired outcomes and processes for ESL/EFL writing instruction and learning.

Some Tradeoffs

Despite these advances, it is important to note some of the associated tradeoffs and costs associated with these advances. First, given the multiple writing tasks that must be generated by test developers and evaluated by human raters, as well as the extended testing time, the TOEFL 2000 writing test is likely to be more expensive to generate, score, and administer than the current TWE test. The fact that the test will likely be more expensive has raised concerns that the test will become more elitist; for example, potential international students who lack funds but who may be eligible for fellowships may not be able to afford the higher-priced tests, especially when other higher-priced admissions tests are also required of

the same examinees. Second, many language teachers around the world lack communicative proficiency and are not prepared to teach from a more integrative, communicative instructional model, although the trend in language teaching and needs in language education are clearly toward such a model (Dickson & Cumming, 1996). This TOEFL 2000 writing framework lays out explicit linkages to this model, thereby improving the potential washback effects of the test. Third, the changes in the test may not be greeted positively by some examinees because of the demands made on their English abilities. For example, examinees who have not had the opportunity to develop their oral and written competencies to the extent that they are realistically prepared for (or close to being prepared for) university studies in an English-speaking educational environment may not perform as well on the TOEFL 2000 test as on the current TOEFL test. Finally, in providing more extensive information on examinees' performance that will be more useful to language teachers, we may be providing more information than the score users actually need.

Thus, while the advances in the TOEFL 2000 writing test warrant moving forward and we can envision ways of addressing the washback and score information concerns, we think that it is important to consider the price at which these advances are gained and the support that score users, language instructors, and examinees are likely to need as the program makes transitions to the next generation of computer-based language testing. These are challenges, nonetheless, that we believe the TOEFL 2000 test can overcome, and indeed should endeavor to, in order to accomplish its primary goals.

References

Alderson, C. (1993). The relationship between grammar and reading in an English for academic purposes test battery. In D. Douglas & C. Chapelle (Eds.), *A new decade in language testing research* (pp. 203-219). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Alderson, C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 116-129.

Allison, D. (1995). Assertions and alternatives: Helping ESL undergraduates extend their choices in academic writing. *Journal of Second Language Writing, 4*, 1-15.

Almond, R. G., & Mislevy, R. J. (1997). *Graphical models and computer adaptive testing* (TOEFL Technical Report No. 14). Princeton, NJ: Educational Testing Service.

Anderson, W., Best, C., Black, A., Hurst, J., Miller, B., & Miller, S. (1990). Cross-curricular underlife: A collaborative report on ways with words. *College Composition and Communication, 41,* 11-36.

Applebee, A. N., Langer, J. A., Mullis, I. V. S., Latham, A. S., & Gentile, C. (1994). *NAEP 1992 writing report card* (National Center for Education Statistics Report No. 23-W01). Washington, DC: US Department of Education.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press. (See chapters 4-5.)

Bachman, L., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Bailey, K. (1999). *Washback in language testing* (TOEFL Monograph Series Report No. 15). Princeton, NJ: Educational Testing Service.

Baker, E. L. (1982). The specification of writing tasks. In A. C. Purves, & S. Takala (Eds.), *An international perspective on the evaluation of written composition* (Evaluation in Education: An International Review Series, Vol. 5, No. 3). Oxford: Pergamon Press.

Baker, E. L., & Quellmalz, E. (1978). *Studies in test design.* Los Angeles, CA: Center for the Study of Evaluation, UCLA.

Bardovi-Harlig, K. (1997). Another piece of the puzzle: The emergence of the present perfect. *Language Learning, 47*, 375-422.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11*, 17-34.

Barzun, J., & Graff, H. J. (1977). *The modern researcher.* New York: Harcourt, Brace College Publishers.

Behrens, L. (1978, Sept.). Writing, reading, and the rest of the faculty: A survey. *English Journal* (Sept.), 54-60.

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

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Erlbaum.

Berkenkotter, C., & Huckin, T. (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Hillsdale, NJ: Erlbaum.

Braine, G. (1995). Writing in the natural sciences and engineering. In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 113-134). Norwood, NJ: Ablex.

Breland, H. (1996). *Writing skill assessment: Problems and prospects.* Princeton, NJ: Educational Testing Service Policy Information Center.

Bridgeman, B., & Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students* (TOEFL Research Report No. 15). Princeton, NJ: Educational Testing Service.

Brossell, G. (1986). Current research and unanswered questions in writing assessment. In K. Greenberg, H. Wierner, & R. Donovan (Eds.), *Writing assessment: Issues and strategies*. New York: Longman.

Brown, J. D. (1991). Do English and ESL faculty rate writing samples differently? *TESOL Quarterly, 25,* 587-603.

Burstein, J. C., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1999). *Computer analysis of essay content for automated score prediction* (TOEFL Monograph Series Report No. 13). Princeton, NJ: Educational Testing Service.

Burstein, J. C., Kukich, K., Wolff, S., Lu, C. (1997). *Final report on automated essay scoring of GMAT essays from the On-line Scoring Network (OSN) October 1997 administration*. Unpublished report. Princeton, NJ: Educational Testing Service.

Butler, F., Eignor, D., Jones, S., McNamara, T., & Suomi, B. (1999). *TOEFL 2000 speaking framework*: *A working paper*. (TOEFL Monograph Series Report No. 20). Princeton, NJ: Educational Testing Service.

Campbell, C. (1990). Writing with others' words: Using background reading texts in academic compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 211-230). New York: Cambridge University Press.

Canale, M., & Swain, M. (1980). Theoretical basis of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1-47.

Carroll, J. B. (1975). *The teaching of French as a foreign language in eight countries*. New York: John Wiley and Sons.

Carson, J., Chase, N., Gibson, S., & Hargrove, M. (1992). Literacy demands of the undergraduate curriculum. *Reading Research and Instruction, 31*, 25-50.

Carson, J., & Leki, I. (Eds.). (1993). *Reading in the composition classroom: Second language perspectives*. Boston: Heinle and Heinle.

Chan, S. (March, 1997). *Methods for assessing and displaying features of coherence in ESL writing*. Paper presented at the Annual Language Testing Research Colloquium, Orlando, FL.

Chapelle, C., Grabe, W., & Berns, M. (1997). *Communicative language proficiency: Definition and implications for TOEFL 2000* (TOEFL Monograph Series Report No. 10). Princeton, NJ: Educational Testing Service.

Chiseri-Strater, E. (1991). *Academic literacies: The public and private discourse of university students*. Portsmouth, NH: Boynton/Cook.

Collis, K., & Biggs, J. (1983). Matriculation, degree structures, and levels of student thinking. *Australian Journal of Education, 27*, 151-163.

Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second-language writing*. New York: Cambridge University Press.

Connor, U. M., & Kramer, M. G. (1995). Writing from sources: Case studies of graduate students in business management. In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 155-182). Norwood, NJ: Ablex.

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

Connors, R., & Lunsford, A. (1988). Frequency of formal errors in current college writing, or Ma and Pa Kettle do research. *College Composition and Communication, 39*, 395-409.

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

Cumming, A. (1995). Fostering writing expertise in ESL composition instruction: Modeling and evaluation. In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 375-397). Norwood, NJ: Ablex.

Cumming, A. (1998). Theoretical perspectives on writing. In W. Grabe (Ed.), *Annual Review of Applied Linguistics, 18*, 61-79. New York: Cambridge University Press.

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

Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 72-93). Clevedon, UK: Multilingual Matters.

Cumming, A., Rebuffot, J., & Ledwell, M. (1989). Reading and summarizing challenging texts in first and second languages. *Reading and Writing, 2*, 201-219.

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

Deckert, G. (1993). Perspectives on plagiarism from ESL students in Hong Kong. *Journal of Second Language Writing, 2*, 131-148.

Denscombe, M., & Robins, L. (1980). Self-assessment and essay writing. *Teaching Sociology, 8*, 63-78.

Dickson, P., Boyce, C., Lee, B., Portal, M., Smith, M., & Kendall, L. (1987). *Assessment of performance unit: Foreign language performance in schools, Report of the 1985 survey of French*. Slough, UK: National Foundation for Educational Research in England and Wales.

Dickson, P., & Cumming, A. (Eds.). (1996). *Profiles of language education in 25 countries*. Slough, England: National Foundation for Educational Research.

Dong, Y. (1996). Learning how to use citations for knowledge transformations: Non-native doctoral students' dissertation writing in science. *Research in the Teaching of English, 30*, 428-457.

Eblen, C. (1983). Writing across the curriculum: A survey of a university faculty's views and classroom practices. *Research on the Teaching of English, 17,* 343-349.

Educational Testing Service. (1996). *TOEFL Test of Written English guide* (4th ed.). Princeton, NJ: Author.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*, 139-155.

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1997). *Computer analysis of the TOEFL Test of Written English (TWE).* Unpublished draft final report. Princeton, NJ: Educational Testing Service.

Freedman, A., & Pringle, I. (1980). Writing in the college years: Some indices of growth. *College Composition and Communication, 31*, 311-322.

Ginther, A., & Grant, L. (1996). *A review of the academic needs of native English-speaking college students in the United States* (TOEFL Monograph Series Report No. 1). Princeton, NJ: Educational Testing Service.

Gorman, T. P., Purves, A. C., & Degenhart, R. E. (Eds.). (1988). *The IEA study of written composition I: The international writing tasks and scoring scales*. Oxford: Pergamon.

Grabe, W., & Kaplan, R. (1996). *Theory and practice of writing: An applied linguistic perspective*. London: Longman.

Grant, L., & Ginther, A. (1995). *TOEFL score user survey report.* Unpublished report. Princeton, NJ: Educational Testing Service.

Hale, G. (1992). *Effects of amount of time allowed on the Test of Written English* (TOEFL Research Report No. 39). Princeton, NJ: Educational Testing Service.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs* (TOEFL Research Report No. 54). Princeton, NJ: Educational Testing Service.

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

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

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

Hayes, R. J. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1-27). Mahwah, NJ: Erlbaum.

Hidi, S., & McLaren, J. (1990). The effect of topic and theme interestingness on the production of school expositions. In H. Mandl, E. DeCorte, N. Bennett, & H. F. Friedrich (Eds.), *Learning and instruction in an international context* (Vol. 2.2, 295-308). Oxford: Pergamon.

Hoetker, J. (1982). Essay examination topics and student writing. *College Composition and Communication, 33*, 377-392.

Homburg, T. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18*, 87-107.

Hughey, J., Wormuth, D., Hartfiel, V. F., & Jacobs, H. (1983). *Teaching ESL composition: Principles and techniques*. Rowley, MA: Newbury House.

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

Hyland, K., & Milton, C. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6*, 183-205.

Hymes, D. H. (1996). *Ethnography, linguistics, narrative inequality*. Bristol, PA: Taylor and Francis, Inc.

IELTS handbook. (1995). Cambridge: University of Cambridge Locale Examination Syndicate.

Intaraprawat, P., & Steffensen, M. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing, 4*, 253-272.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4*, 51-69.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (1999). *TOEFL 2000 framework: A working paper*. (TOEFL Monograph Series Report No. 16). Princeton, NJ: Educational Testing Service.

Janopoulos, M. (1992). University faculty tolerance of NS and NNS writing errors: A comparison. *Journal of Second Language Writing, 1*, 109-121.

Johns, A. M. (1985). Summary protocols of "underprepared" and "adept" university students: Replications and distortions of the original. *Language Learning, 35*, 495-517.

Johns, A. M. (1990). L1 composition theories: Implications for developing theories of L2 composition. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 24-36). New York: Cambridge University Press.

Johns, A., & Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied Linguistics, 11*, 253-271.

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

Krapels, A. R. (1990). An overview of second language writing process research. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 37-56). New York: Cambridge University Press.

Kroll, B. (1979). A survey of writing needs of foreign and American college freshmen. *ELT Journal, 33*, 219-226.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 140-154). New York: Cambridge University Press.

Langer, J. A., & Applebee, A. N. (1987). *How writing shapes thinking: A study of teaching and learning*. Urbana, IL: National Council of Teachers of English.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in written production. *Applied Linguistics, 16*, 307-322.

Leki, I., & Carson, J. (1994). Students perspectives of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly, 28*, 81-101.

Master, P. (1986). *A cross-linguistic interlanguage analysis of the acquisition of the English article system.* Unpublished doctoral dissertation, University of California, Los Angeles.

McCarthy, L. P. (1987). A stranger in strange lands: A college student writing across the curriculum. *Research in the Teaching of English, 21,* 233-265.

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

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 12-23.

Messick. S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241-256.

Mislevy, R. J. (1994a). *Probability-based inference in cognitive diagnosis* (ETS Research Report No. 94-3-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1994b). *Test theory and language learning assessment.* Paper presented at the Center for Advancement of Language Learning, 1994 Language Aptitude Invitational Symposium, Arlington, VA.

Mohan, B. (1986). *Language and content.* Reading, MA: Addison-Wesley.

Nelson, J. (1990). This was an easy assignment: Examining how students interpret academic writing tasks. *Research in the Teaching of English, 24,* 362-395.

Olson, D. R. (1996). *The world on paper: The conceptual and cognitive implications of writing and reading.* Cambridge: Cambridge University Press.

Ostler, S. E. (1980). A survey of academic needs for advanced ESL. *TESOL Quarterly, 14,* 489-502.

Paltridge, B. (1994). Genre analysis and the identification of textual boundaries. *Applied Linguistics, 15*(3), 288-299.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly, 14,* 61-69.

Pierce, B. N. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly, 26,* 665-691.

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

Powers, D. E., & Fowles, M. E. (1996). Effects of applying different time limits to a proposed GRE writing test. *Journal of Educational Measurement, 33,* 433-452.

Powers, D. E., & Fowles, M. E. (1998). Effects of pre-examinations disclosure of essay topics. *Applied Measurement in Education, 11*, 139-158.

Powers, D. E., Fowles, M. E., & Boyles, K. (1996). *Validating a writing test for graduate admissions* (GRE Research Report No. 93-26bP). Princeton, NJ: Educational Testing Service.

Purves, A. (1992). Reflection on research and assessment in written composition. *Research in the Teaching of English, 26*, 108-122.

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

Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing, 1*, 79-107.

Reynolds, D. (1995). Repetition in nonnative speaker writing: More than quantity. *Studies in Second Language Acquisition, 17*, 185-209.

Rifkin, B., & Roberts, F. (1995). Error gravity: A critical review of research design. *Language Learning, 45*, 511-537.

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

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

Santos, T. (1992). Ideology in composition: L1 and L2. *Journal of Second Language Writing, 1,* 1-15.

Sarig, G. (1993). Composing a study-summary: A reading/writing encounter. In J. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp. 161-182). Boston, MA: Heinle and Heinle.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning, 46*, 137-174.

Savignon, S. (1983). *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley.

Sherwood, R. (1977). A survey of undergraduate reading and writing needs. *College Composition and Communication, 28,* 145-149.

Silva, T. (1990). Second language composition instruction: Developments, issues, and directions in ESL. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 11-23). New York: Cambridge University Press.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly, 27,* 657-677.

Smith, W., Hull, G., Land, R., Moore, M., Ball, C., Dunham, D., Hickey, L., & Ruzich, C. (1985). Some effects on varying the structure of a topic on college students' writing. *Written Communication, 2,* 73-89.

Sperling, M. (1996). Revisiting the writing-speaking connection: Challenges for research on writing and writing instruction. *Review of Educational Research, 66,* 53-86.

Swales, J. (1990). *Genre analysis*. New York: Cambridge University Press.

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

Taylor, C. (1993). *Report of TOEFL score users focus groups.* Unpublished report. Princeton, NJ: Educational Testing Service.

Taylor, C., Eignor, D., Schedl, M., & DeVincenzi, F. (1995, March). *TOEFL 2000: A project overview and status report*. Paper presented at the annual meeting of TESOL, Long Beach, CA.

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

The National Assessment Governing Board. (n.d.). *Writing framework and specifications for the 1998 National Assessment of Educational Progress.* Washington, DC: Author.

Tobias, S. (1994). Interest, prior knowledge, and learning. *Review of Educational Research, 64,* 37-54.

Van der Geest, T. (1996). Studying "real-life" writing processes: A proposal and an example. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 309-415). Mahwah, NJ: Erlbaum.

Vann, R., Meyer, D., & Lorenz, F. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly, 18,* 427-440.

Walvoord, B., & McCarthy, L. (1990). *Thinking and writing in college.* Urbana, IL: NCTE.

Waters, A. (1996). *A review of research into needs in English for academic purposes of relevance to the North American higher education context* (TOEFL Monograph Series Report No. 6). Princeton, NJ: Educational Testing Service.

Weir, C. J. (1984). The Associated Examining Board's Test in English for Academic Purposes (TEAP). In R. Williams, J. Swales, & J. Kirkman (Eds.), *Common ground: Shared interests in ESP and communication studies* (pp. 145-158). (ELT Documents No. 117). London: The British Council/Pergamon.

Whalen, K., & Menard, N. (1995). L1 and L2 writers' strategic and linguistic knowledge: A model of multiple-level discourse processing. *Language Learning, 45*, 381-418.

White, E. M. (1988). *Teaching and assessing writing*. San Francisco: Jossey-Bass.

White, E. M. (1995). An apologia for the timed impromptu essay test. *College Composition and Communication, 46*, 30-45.

Wikborg, E. (1990). Types of coherence breaks in Swedish student writing: Misleading paragraph division. In U. Connor & A. Johns (Eds.), *Coherence in writing research and pedagogical perspectives* (pp. 131-149). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Zhang, S. (1995). Semantic differentiation in the acquisition of English as a second language. *Language Learning, 45*, 225-249.

Appendix A: General Standards for Good Writing

The standards for good writing described in this section are ones that we feel apply to the writing expected by the vast majority of institutions that make use of TOEFL test scores. We recognize, however, that there are different genres and rhetorical traditions within the English-medium tradition and a far greater variety when other language groups and cultures are examined. Grabe and Kaplan (1996) present a summary of issues related to what is often termed "contrastive rhetoric." They point out (p. 198) that "there are rhetorical differences in the written discourses of various languages, and that those differences need to be brought to consciousness before a writer can begin to understand what he or she must do in order to write in a more native-like manner (or in a manner acceptable to native speakers of the target language)." Our attempt here is to describe qualities of writing valued in academic contexts in English-speaking North America. In Section 3 we use some of these notions to discuss a possible model of the reader who will be rating the writing tasks that ultimately will constitute the TOEFL 2000 test.

One acid test of good writing, as applied by Barzun and Graff (1977), is "can another trained mind, not expert in my subject, understand what I am saying?" (p. 27). *Any* interested reader should be able to grasp the writer's message with "only a normal amount of effort and attention (p. 27)." For Barzun and Graff (1977), "the expression [of knowledge] is the knowledge. What is not properly presented is simply not present—and its purely potential existence is useless" (p. x).

Denscombe and Robins (1980) discuss deficiencies in persuasive writing in terms of topics being introduced in a "strange and inappropriate order," with no apparent logical order, thus inhibiting persuasiveness and leading the reader to become "confused and dubious" about the writer and his/her understanding of the subject. A clear argument requires an initial statement of the theme or the stance to be taken by the writer, as well as a clear indication of how the significance of subsequent points relies on previous ones. Not only should writers avoid straying from their subject, they should also strive to ensure that the material discussed is relevant and that its relevance is explained. Writers may also fail to back up assertions or assume, when referring to statistics, charts, and tables, that the data "speak for themselves," expecting their readers to accept their pronouncements as an act of faith. In some kinds of writing, persuasiveness is accomplished by citing other sources to support one's points.

In short, good writing can be defined in terms of the general overall impression that the writer makes on his or her reader. In this view, anything that interferes with the reader's understanding is bad writing; a good writer uses skills and strategies that help to convey meaning and avoids the use of practices that obscure meaning. It is important to recognize, though, that the criteria and standards for such judgments vary with the context, purposes, and expectations for writing, as well as with the interpersonal roles that writers and readers may share. A brief e-mail message to set an appointment with a friend and a technical report submitted for a university course assignment, for example, are likely to differ greatly in the qualities that determine explicitness, persuasiveness, and perceived satisfaction.

Standards and expectations as to what constitutes good writing may vary according to the particular context in which the writer finds himself/herself. As White (1995) has noted, the expectations for undergraduate students are typically much lower than for graduate students. For the former, being required to engage in writing activities is often in order to foster inquiry and reflection and analysis of complex arguments as an entry to the genre of a discipline. For graduate students, expectations may involve using sources effectively to support ideas, analyzing and refining complex arguments, exhibiting

considerable depth of understanding of a subject, and engaging in the discourse in one's field. Whereas undergraduate students are asked primarily to master a body of knowledge, graduate students are often asked to go further—to make an original contribution to their field. Moreover, graduate students are more often expected to have mastered many of the genres of writing that undergraduate students are still acquiring as they enter specific disciplines.

Generally, most of the common notions about writing stress the ability to organize thoughts, to say something that engages the reader, to exhibit depth in some sense, to remain on-task and to-the-point, and to provide a sense of closure. Variety, a clear theme or thesis, connections among and organization of ideas, detail and elaboration when necessary, and fluency/smoothness are other oft-mentioned properties of good writing. All of these notions are consistent, however, with the more general view that good writing can be defined in terms of the general overall impression that writers make on their readers. This criterion—of overall impression—is in fact the one that is applied in traditional holistic evaluations of writing skill. However, specific criteria to make this central criterion operational relate mostly to formal genres of extended writing, such as essays or reports. As such, they may not apply as directly to other, less formal or extensive writing tasks, such as writing letters, completing forms, or doing textbook exercises (which are also relatively common to, and important for, academic studies).

Appendix B: Example of Independent Invention Essay and Response

Independent Invention Essay Stimulus

Essay: Supporters of technology say that it solves problems and makes life better. Opponents argue that technology creates new problems that may threaten or damage the quality of life. Using one or two examples, discuss these two positions. Which view of technology do you support? Why?

Prototypical Response (sample TWE essay that received a score of 5.0)

Technology by definition refers to the improvement of the technical know how and advancement of machinery to improve the working systems in the human society. In a way this looks a very good idea in that mans' work would be made much faster and less laborious. Machines which are the main implements of technology have a major advantage to mans' ways of life. Take for example an aeroplane, which being a product of advance in technology has made all corners of the earth look like they are only centimetres apart. It has made the means of communication which prior to its development was very difficult much easier and less risky. Travelling to many parts of the world which are very many miles apart now only takes a few hours or days whereas this used to take days or even months.

On the other hand technology has created a number of new hazards to the health of the societies. The machines make life easy but also expose people to new problems. In the example considered above transportation has become easier by planes but these planes also expose people to accidents which have become so numerous and claim many lives daily. As we all know that a majority of these machines use fuel and that to use the fuel it has to burn there are new products which are introduced into our environment. These new products include gases from automobiles which pollute the air we breathe. These gases expose us to lung diseases, cancers and number of new ailments which have not yet been fully explored.

In conclusion I think that although advances in technology may seem favourable there are alot of hazards which it introduces into our ways of life.

Appendix C: Example of Interdependent Text-based Stimulus and Response

Interdependent Text-based Stimulus

What Enabled Early Humans to Control the Use of Fire

What was it that enabled early humans to control the use of fire; first to keep a fire going for an extended length of time and then to be successful in passing on this ability from generation to generation? In order to answer this question, it may be useful to distinguish between the physical, mental, and social preconditions that were necessary. No doubt such physical features as erect posture and the concomitant aptitude for carrying objects in the hand and manipulating them were essential. Even before humans could make fires themselves, one of the advantages which they (and possibly other primates as well) had over other animals was that they were able to handle sticks with which they could rummage in the smoldering fire without getting burned. After a forest fire they were able to search through the ashes for food and probably noticed that they might prolong the fire's burning by throwing branches on it. Even more important, however, was the capacity to pick up burning matter and transport it to a place where it could not be extinguished by rain or wind.

But this was clearly not just a matter of physical advantages of early humans, of erect posture and having the hands free to carry something else. Fetching branches for a fire implies that the individuals concerned thought about what they were doing, and knew why they were doing it. Keeping a fire going implies foresight and care. Wood had to be gathered, and perhaps even stored during wet periods. Such activities did not come naturally to early humans; they required learning and discipline. Especially when humans began to collect fuel over larger distances, they devoted part of their energy to maintaining something outside themselves, something beyond their own immediate needs. This is not to say that they were acting "unselfishly." Tending the fire was a form of "deferred gratification" or putting off the satisfaction of immediate needs in planning for future needs, like that which was later to become an essential ingredient in agriculture and livestock-raising. Unlike superficially similar complex activities such as nest-building by birds, it was not genetically determined but had to be learned.

Reading Comprehension Question

Which of the following is the main topic of the passage?

The positive effects of forest fires on early humans.

Early indications of superior human intelligence.

Characteristics that made it possible for early humans to control fire.

Environmental conditions that threatened the survival of early humans.

Writing Instructions: Based on the reading passage, what were the requirements for the successful use of fire among early humans?

_____

_____

_____

_____

_____

_____

Prototypical Response

There were three main requirements for the successful use of fire among early humans. The first was the physical ability to manage fire (e.g., the size, dexterity, and coordination of human beings). The second was the intellectual ability to learn, remember, and teach the steps necessary to control fire. And the third was the social-psychological ability to think ahead, plan for the future (e.g., by collecting and storing wood), and delay immediate gratification for the sake of more distant benefits. It took all three of these capacities for humankind to advance in its living conditions through the use of fire.

Appendix D: Example of Interdependent Situation-based Stimulus and Response

Interdependent Situation-based Stimulus

Professor Baker to class: "Now that we've read and discussed this story, I'd like you to write an essay on it for our next class. You have two choices for the topic. The first one is to present your interpretation of the reasons the main character decides to move to California at the end of the story. The second is to analyze one significant choice made by any character in the story."

The next day: A study group meets to discuss the assignment.

Ken: Which topic are you going to write on?

Sue: The second. I think it's easier because you can choose any character in the story. What about you?

Ken: I don't know yet. I have to re-read the story before I decide. Lisa, what are you going to do?

Lisa: Well, I want to write about the main character, but I'm not sure what Professor Baker means when he says 'present your interpretation.'

Sue: I think he means analyze why the character decided to move. I mean, if you choose the second topic, you have to analyze, so you probably have to in the first one, too.

Ken: Maybe, but I think he wants you to give your opinion.

Lisa: If I just have to give my opinion, that's easy. But if I have to analyze it, I don't think I can because I don't really know why she moved. It was a surprise ending.

Sue: Why don't you write Professor Baker a note and ask him?

Lisa: Good idea! I think I will because that's the topic I want to do.

Writing Instruction: Write a note to Professor Baker in which you identify and explain the problem and then make a request.

Dear Professor Baker,

_____
_____
_____
_____

Prototypical Response

Dear Professor Baker,

This is _____ in your writing class, and I have a question regarding the first essay topic that we could write about. I don't quite understand what is meant by the word "interpretation." Do you want opinion, analysis, or the character's motive in moving to California? I would appreciate it if you could Email me as soon as possible. Thank you for your time.

Appendix E: TOEFL 2000 Writing Framework Variables

A. Variables that delineate the domain, define features for constructing tasks, control for test assembly, and characterize aspects of task difficulty

- A1 Task stimuli
  - A1.1 Independent invention essay task
    - (a) Identify and explain a problem and offer a solution
    - (b) Take a position on an issue, elaborate it, and/or justify it
    - (c) Perhaps also categorize, define, enumerate, compare/contrast
  - A1.2 Interdependent text-based task
    - (a) Identify, select, and manipulate relevant information from text or texts
  - A1.3 Interdependent situation-based task
    - (a) Identify a problem, explain or elaborate on it, and propose or request a solution
    - (b) Synthesize information and present, summarize, or report on it
- A2 Response format
  - A2.1 Mode of response
    - (a) Word-processed
    - (b) Handwritten
- A3 Type of response
  - A3.1 Elaborated responses (e.g., essays)
  - A3.2 Brief responses (e.g., synthesis, summary, note)

✓* A4 Rhetorical function specified in the task

- ✓ A4.1 Categorize key features and/or analyze and describe relations between them
- ✓ A4.2 Identify a problem and analyze it and/or propose a solution to it
- ✓ A4.3 State a position, elaborate it, and/or justify it

These rhetorical functions include related concepts:

- ✓ A4.4 Define
- ✓ A4.5 Enumerate
- ✓ A4.6 Compare/contrast

✓ A5 Topic Characteristics

- ✓ A5.1 Academic contexts and content
- ✓ A5.2 Extent of required topic-specific background knowledge and personal experience of examinees
- ✓ A5.3 Emotional appeal or interestingness of topic to examinees
- ✓ A5.4 Extent of options or choices available to examinees in their writing
- ✓ A5.5 Extent of topic abstractness or specificity
- ✓ A5.6 Differences in examinees' cultural or geographical backgrounds
- ✓ A5.7 Cognitive and experiential demands of topic
- ✓ A5.8 Information load of topic
- ✓ A5.9 Time elements
- ✓ A5.10 Length of response required
- ✓ A5.11 Access to source material(s)

* = Task model variables hypothesized to characterize aspects of task difficulty.

B. Variables that characterize performance or individual responses (i.e., scoring rubrics and rules for scoring)

- B1 Scoring Rubrics
  - B1.1 Discourse
    - (a) Organization
    - (b) Coherence
    - (c) Progression
    - (d) Development, specificity, and quality of information and ideas
    - (e) For integrated tasks, ability to integrate/summarize/report
  - B1.2 Language use
    - (a) Appropriacy and range of vocabulary and idiom used
    - (b) Appropriacy and effectiveness of language used to manage discourse connections
      - (i) Rhetorical functions, illocutionary markers
      - (ii) Paraphrasing, restating, citing, quoting source material
    - (c) Accuracy, appropriacy, and range of phrase-level syntax and morphological markers
      - (i) Subordinate clauses, participle phrases
      - (ii) Grammatical features such as articles, tense, adverbials
    - (d) Spelling, punctuation, other orthographic or typographic conventions

Appendix F: Software and Technology Issues

This appendix briefly discusses software and technology issues related to: 1) the process of writing, 2) the capture and transmission of essays and other written responses, and 3) the scoring of writing and the possibilities for providing feedback to examinees.

The Process of Writing

One of the goals of the TOEFL 2000 writing assessment stated above is to facilitate a variety of writing processes. Toward this end, various software and hardware options might be explored and research projects developed to investigate the extent to which such options might contribute to or detract from construct validity. In terms of prewriting, in addition to or instead of "scrap paper" for outlining, methods might be investigated to allow examinees to create electronic notes or even to record thoughts orally (for later playback) as they are actually writing or typing their texts. Research into the facilitative or deleterious effects of word-processor features such as spell checkers and thesaurus use in essay testing might be carried out for TOEFL test populations as well. In a future platform, it would not be inconceivable to partner with one or more software concerns to allow one standard word-processing package or a choice of well-known industry standards.

Capturing and Transmitting Essays

If handwritten entry is permissible, investigations might well be conducted into electronic tablets that would allow capture of handwriting for direct electronic transmission to a rating center. The technological challenge would be to simulate for examinees means of erasures and corrections to text in a fashion that would feel as comfortable as or more comfortable than paper-and-pencil composition. Such technological capability would vastly improve on current anticipated procedures for handling handwritten essays: either imaging them on-site and transmitting or mailing the essay answer sheets back to Princeton and then imaging them for distribution to a remote reader network of scorers. In both current scenarios, major issues remain for matching answer sheets with examinee and topic/task numbers. Such difficulties would be avoided by on-line electronic capture of handwriting.

Scoring and Feedback

Nowhere does software innovation promise more added value than in the areas of scoring and feedback. A number of suggestions have been made throughout the body of this paper to investigate aspects of language structure, linguistic accuracy, and mechanics of writing that may be associated with the development of writing skills and the evaluation of writing.

Natural language processing (NLP) tools could be of immense assistance in the formulation and validation of the linguistic aspects of score scales. Making use of word frequency lists, corpuses, grammatical analysis, and pretest essays, NLP analyses could inform efforts to understand topic and task comparability in a quantitative and comparative fashion. Efforts to understand topic and task comparability would be enhanced even further to the extent that NLP analyses might be shown to model rhetorical features such as development of a topic or subtopics, unity, and coherence. Such analyses of pretest data then might help to make final decisions to use potential items and even to contribute to the calibration of those items.

With respect to the scoring process itself, NLP analyses can provide actual counts of linguistic and other elements that could be used in conjunction with holistic ratings of language use or possibly as independent, automatic ratings if they can be shown to give reliable and valid assessment. And for constructed response tasks that involve summary writing, NLP could very easily provide a measure of verbatim use of language from the source text, a standardized measure for every examinee that could be useful in gauging the degree to which examinees are unable to create paraphrase and summary.

Even if NLP is not used for essay or constructed response scoring itself, NLP procedures could be used to monitor and evaluate rater behavior. This might be accomplished through predictive models of scores and comparing such models with how closely individual raters conform. (Analysis of discrepancies between reader scores and predicted scores on individual papers might point out failings on the part of particular raters or point out deficiencies in the NLP scoring model). Alternatively, by comparing rater behavior with particular NLP analyses or counts—total word count perhaps being the simplest exemplar—NLP might help determine if any individual reader's scoring is unduly influenced by particular features that NLP can capture.

Finally, NLP analyses could be used to provide diagnostic feedback to examinees. Measures of vocabulary complexity, of number of words not found in a dictionary (possibly indications of misspellings), of readability levels—these and more would be useful bits of information to accompany actual scores. Again, to the extent that NLP analyses can be shown to reliably indicate higher level aspects of evaluation, feedback on rhetorical qualities of examinees' writing might also be given. What would be useful overall in decisions to use and how to use NLP techniques and outcomes would be a theoretical framework for NLP or some sort of taxonomy of NLP approaches so that the TOEFL 2000 project and other projects could benefit from and incorporate a structured and grounded view of this.

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
E-mail: toefl@ets.org
Web site: http://www.toefl.org

58713-005251 • Y40M.750 • 253719 • Printed in U.S.A.