TOEFL®

Monograph Series

MS - 5
MARCH 1997

TOEFL 2000 — Writing: Composition, Community, and Assessment

Liz Hamp-Lyons

Barbara Kroll

Educational Testing Service

TOEFL 2000 —

WRITING: COMPOSITION, COMMUNITY, AND ASSESSMENT

Liz Hamp-Lyons and Barbara Kroll

Educational Testing Service
Princeton, New Jersey
RM-96-5

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

Copyright © 1997 by Educational Testing Service. All rights reserved.

No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both U.S. and international copyright laws.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, TOEFL, the TOEFL logo, TSE, and TWE are registered trademarks of Educational Testing Service.

Foreward

The TOEFL® Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language program 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 were invited 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 ETS will evolve into the 21st century. As a first step in the evolution of TOEFL language testing, the TOEFL program recently revised the Test of Spoken English (TSE®) test and announced plans to introduce a TOEFL computer-based test (TOEFL CBT) in 1998. The revised TSE, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The TOEFL CBT will take 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

It is expected that the TOEFL 2000 efforts will continue to produce a set of improved language tests that recognize the dynamic, evolutionary nature of assessment practices and that promote responsiveness to test user needs. As future papers and projects are completed, monographs will continue to be released to the public in this new TOEFL research publication series.

TOEFL Program Office
Educational Testing Service

Abstract

Commissioned by the TOEFL program in 1993, this monograph explores the salient issues of an approach to assessing writing in the context of the TOEFL test and in light of what is currently known and believed about the acquisition and assessment of writing. It encompasses the Committee of Examiners' model of communicative competence (1993) as well as the accepted paradigms for the teaching and learning of writing, and suggests an attention to writing as "discourse competence," that is, writing as an act that takes place within a context, accomplishes a particular purpose, and is appropriately shaped for its intended audience.

The monograph considers these questions in a wide-ranging exploration: What skills are needed to succeed in an academic context? Are these skills the same for undergraduate and graduate students? Are the performance expectations the same for these two groups? How can writing be elicited that is true to the nature of "real" writing in the academy? Whose expectations should be privileged: English faculty, subject-area faculty teaching at lower academic levels, or subject-area faculty teaching at advanced, specialized levels? How can we train readers to judge this writing appropriately? What are the appropriate bases for judgments of writing of these kinds? How do we ensure that test takers, whatever their backgrounds, are equally challenged?

The monograph describes a variety of potential approaches to writing assessment and considers how these might be applied to academic writing in the TOEFL 2000 context. It pays particular attention to the attempt to describe the test taker population of TOEFL 2000, pointing out how little is known about test takers generally and suggesting the need for studies of writers and their reactions to test prompts and conditions. It then considers prompt development, scoring procedures, score reporting, and score use in some detail, drawing attention to the special problem of the limited time usually available in test writing conditions.

Finally, the monograph closes with a consideration of costs, practicability, and washback from the test, and makes some key recommendations for writing test development for TOEFL 2000.

Acknowledgments

We would like to acknowledge the support of our fellow members of the Test of Written English (TWE®) Committee, and especially of Carol Taylor, first as director for the TWE program and then as TOEFL 2000 project coordinator, and of Robbie Kantor, the current TWE program director.

Table of Contents

Introduction

Of all the language skills, writing is the one that can affect a student's college career the most. Most of the gatekeeping practices in the universities and colleges of English-speaking countries require the production of written text. Further, most of these practices involve dual expectations of competence in the subject matter (whatever it may be) and of competence in the medium through which mastery of that subject matter is shown — writing. In most colleges, writing is a core competency, a keystone in the foundation on which intellectual and disciplinary development is based. In part because of its critical role in examinations and the other formal hurdles placed before college students, a great deal of writing is expected of college students in less formal and critical contexts: writing lab reports, analytic responses to texts, creative writing, and research papers. These and a range of other genres are expected throughout college (Hale, Taylor, Bridgeman, Carson, Kroll, & Kantor, 1996).

But writing plays other roles in a student's college life; note-taking from lectures and text materials, summarizing, and, increasingly, keeping learning logs and journals all place heavy demands on students' written competence, while at the same time playing a vital part in their creating their own understanding of the ideas they encounter (Walvoord, 1986). While good listening and reading skills are essential to students' reception of the knowledge the academy has to offer, speaking and writing are essential to students' knowledge integration and production. For nonnative speakers (NNS), weak writing skills can not only lead directly to failure in formal assignments and examinations, but indirectly to lost opportunities to engage in text-making as knowledge-making (Herrington & Moran, 1992; Langer & Applebee, 1986). In looking toward TOEFL 2000, it becomes imperative not merely to include an evaluation of writing within the new test design, but to ensure that the kind(s) of writing tested, the purposes stated, and the audiences specified reflect as closely as possible the communicative demands of U.S. and Canadian colleges and universities at a range of levels and across disciplines.

The number of foreign visa-holding candidates seeking admissions to North American universities shows no sign of decreasing, and the need to ensure a minimum proficiency in English among entering students matriculating at the undergraduate or graduate level will continue to exist. As the main provider of these testing services, ETS will continue to provide a major service to college and university admissions while moving toward a test for the new century that will reflect all we have come to know both about academic language performance and about language assessment. For the purposes of this discussion, we will assume that TOEFL 2000 will, like TOEFL, be an academic entrance requirement playing a role in the admissions decisions at large numbers of North American institutions of higher education.

We begin this paper by addressing the notion of communicative competence as it might apply to assessing communicative competence in writing. Pointing out the origins and associations of present thinking in the fields of composition and writing assessment, we suggest that a more appropriate theoretical perspective from which to envision a writing test is the notion of discourse competence. We go on to review current research in writing assessment by considering the following as the major building blocks in designing the writing component for the TOEFL 2000 test: the construct (writing skills), the context (assessment variations), and the instrumentation (the writing test/task and the scoring procedures). We conclude with some recommendations for further investigation in the area of writing assessment for second-language writers.

From Communicative Competence to Theory and Models in Composition

To begin to discuss how best to assess the writing proficiency of large numbers of test candidates, we must first consider the nature of writing itself, and how it might fit into a larger model of the communicative competencies that the TOEFL 2000 wishes to measure.

As proposed by Hymes (1972), communicative competence served as a richer view of linguistic competence than the model first proposed by Chomsky (1957, 1966), which had previously served to change the way language itself was seen. To consider a speaker's communicative competence meant to attend to the entire package of a particular speaker's linguistic performance rather than focusing on language competence and performance (language and parole) at and below the sentence. This was one approach to judging the speaker's ability to present herself or himself as a member of a particular discourse community. As applied to the second-language arena, the notion of communicative competence served to revolutionize pedagogy in second-language classrooms: the purpose of second-language instruction became the turning out of speakers/writers who could successfully participate in a wide range of linguistic events, and the previous emphasis on rote learning and grammatical accuracy was given lower priority.

In the 1970s this model of communicative competence was generally accepted in English as a Second Language (ESL) and English as a Foreign Language (EFL) teaching programs in English-speaking countries (although it must be remembered that traditional methods of teaching English did and do still predominate in many non-English-speaking countries). Johnson (1983) modified Hymes' model of communicative competence for the development of materials to teach writing communicatively (these were published as Johnson, 1981). Johnson illustrates why a strict communicative (notional or functional) syllabus fails to prepare students adequately for the nonstereotypical nature of written communicative events. He prefers instead an approach that prepares students by presenting examples of written utterances at the level of discourse and guiding them to explore the relationships between the discourse's constituent utterances, their contexts, and their intents. In this approach, reading comprehension activities figure strongly, leading naturally into writing practice.

In the past 10 or 15 years, however, ESL composition theorists and practitioners have generally abandoned notions of communicative competence in favor of models and approaches from first-language composition, and the teaching of writing to ESL students in colleges and universities in North America has been influenced not by sociolinguistics, the theoretical field from which communicative competence arises, but by rhetoric and composition studies. In fact, while ESL/EFL teacher training is rooted in various subfields of linguistics, psychology, and education, ESL/EFL writing teachers who practice at the tertiary level find it essential to become familiar with work in the field of composition studies, a subfield of English studies. Raimes (1983) suggested that "we have stressed the ESL part of ESL composition at the expense of the composition part." And in the early 1980s second-language composition teacher-researchers began to pay attention to "the composition part," asking the same kinds of questions about the effectiveness of different teaching strategies on student learning as were being asked in first-language composition research. For instance, Zamel (1983) used case study research methodology much like that of the ground-breaking studies in L1 composition by Perl (1979), Sommers (1980), and others.

She found that focusing on grammatical accuracy and conformity to rhetorical conventions led students to make only surface revisions that limited their growth as writers. Going beyond L2 writing instruction as teaching grammar, paragraph models, and error correction and beyond Johnson's L2 writing instruction as communicative discourse where "discourse" is teaching and practicing rhetorical forms into which meanings are to be poured, Raimes and Zamel spearheaded the concern in ESL composition studies with writing as the expression of ideas, as a conveyance of meaning. But they went further, agreeing with Berthoff (1978) that writing is both knowledge creation and form-finding — both a way to find out what we want to say and a way to discover the appropriate discourse vehicle for what we have to say. In this they blended expressivist concerns with cognitive approaches most strongly associated with the work of Hayes and Flower (1983) and Flower and Hayes (1981, 1984). Johns (1990) says:

> The influence of the process approaches, especially of cognitive views, upon modern ESL writing classrooms cannot be exaggerated. In most classrooms, teachers prepare students to write through invention and other prewriting activities, encourage several drafts of a paper, require paper revision at the macro levels, generally through group work, and delay the student fixation with the correction of sentence-level errors until the final editing stage (p. 27).

Johns (1990) reminds us that Flower and Hayes (1981, 1984) were concerned with thinking (with *cognitive* processes), and that the elements of the writing process so common in ESL writing classrooms today are elements that Flower and Hayes found in good writers; that is, in writers who "not only have a large repertoire of powerful strategies, but they have sufficient self-awareness of their own process to draw on these alternative techniques as they need them. In other words, they guide their own writing process" (Flower, 1985, p. 370: quoted in Johns, 1990, p. 27).

While teachers of second-language composition must know not only about language but about rhetoric and allied studies as they apply both to L1 and L2 writers, we must remember that we cannot presume that the act of writing in one's first language is the same as the act of writing in one's second language. As Kroll (1990) pointed out:

> What teachers need is an understanding of all facets of this complex field of writing, and then to filter that understanding through a prism that can reflect how the factor of using a nonnative code affects second language writing performance (p. 2).

In an important overview, Leki (1992) tells us that "when research on L2 composing processes finally began in the early 1980s, they revealed basic similarities between L1 and L2 writers, concluding that as far as composing processes were concerned, the distinction to be made was not between L1 and L2 writers but between experienced and inexperienced writers." Findings like these led researchers to endorse the imitation of L1 writing classroom practices in ESL writing classrooms. Students who will take the TOEFL 2000 test will presumably be fully competent speakers and highly literate writers in their native language(s). Research [for example, by Cumming (1989) and Zamel (1983)] has shown that ESL/EFL students who are expert writers in their own languages are able to use the same writerly strategies they have learned in their L1 as they compose in English, their L2. However, Raimes (1985) found that while this was generally true, the effectiveness of students' use of their writing strategies was not always as

great in the L2 as in the L1. We must keep in mind Leki's caution that there are differences too between L1 and L2 writing processes, and that these differences remain relatively unexplored (p. 79).

But attention to composing processes has been only one of the perspectives on second-language writing research and teaching in the past 10 to 15 years. Silva (1990) reminds us that critics of the process approach in ESL writing instruction have questioned whether it realistically prepares students for academic work, and have endorsed instead an English for Academic Purposes (EAP) approach. The EAP approach focuses on academic discourse genres, on the range and nature of academic writing tasks, and on socializing students into the academic discourse community. In this, it has the same genesis as social constructionist views in first-language rhetoric/composition studies. A social constructionist view of writing, growing from the work of Kuhn (1970) and exemplified in the work of Becher (1987), Bruffee (1986), and Herrington (1986), sees college students as apprentice members of the academic discourse community, with writing needs — and language needs generally — that depend on the specific disciplinary communities they enter. Johns (1990) points out that in this approach, rather than being *creators*, as in the process approach, writers are *socially constructed.* The language, focus, and form of their texts are dependent on the community in which and for which they are writing; they are acted upon rather than actors. In the second-language arena, this approach is especially associated with the work of Swales (1990), Johns (1991a, 1991b), and Horowitz (1986). The current study by Hale et al. (1996) exemplifies a social constructionist approach to the exploration of the writing needs of ESL college students.

While the two approaches described above — the cognitive process approach and the social constructionist model — have the greatest currency for those of us in second-language writing, in the field of rhetoric and composition as a whole, other models make more complex distinctions. For instance, Knoblauch (1988) proposes four rhetorical traditions: ontological (Aristotelian), objectivist (Locke), expressionist (Kant and, we assume, Britton), and sociological (Bahktin). The ontological tradition sees writing through its formal linguistic properties and sees it as critical in transmitting knowledge but not in creating knowledge; such a view is behind current traditional approaches to the teaching of writing, and until recently has explicitly informed most approaches to the assessment of second-language writing. The objectivist framework emphasizes the role of writing in creating knowledge, seeing knowledge as dependent on discourse, and might be seen as related to the cognitive model of Flower and Hayes. The expressionist tradition situates knowledge in the human imagination and sees writing as a conduit for this. Expressionist rhetoric has its origins in Britton's influential functional model (Britton, Burgess, Martin, McLeod, & Rosen, 1975), which introduced such terms as "transactional" writing and "expressive" writing. In the Britton et al. model, writers move from inner, personal writing, or self-expression, to three forms of public discourse (transactional, expressive, and poetic writing), effectively creating these forms of discourse for themselves. This expressivist view of writing remains common today. Finally, the sociological (dialogic) view sees writing, like all language, as a social practice growing from cultural and material processes, that can be seen as related to the social constructionist model discussed above. The model proposed by Berlin (1982) describes three related rhetorics: cognitive, expressionist, and social epistemic. Social-epistemic rhetoric, as exemplified in the first-language field by the work of Young, Becker, and Pike (1970); Faigley (1986); and Bartholomae (1985), and in applied linguistics by the work of Pennycook (1989) and Auerbach (1986), sees rhetoric as a political act with language as the agency of mediation (Berlin, 1982, p. 488). The perspective of social-epistemic rhetoric has particular implications for the development of a new TOEFL test, for its proponents are very aware of the sociopolitical dimensions of the test design and

administration decisions. This is exemplified in the criticisms of the current TWE test and the TOEFL program policies and structures by a key second-language composition figure, Raimes (1990), and her colleague Greenberg (1986), a key figure in first-language writing assessment.

Each of the models we have discussed presents alternative ways of viewing the problems and potentials writers face as they approach a writing event, and alternative ways of looking at that event. Clearly, there is no generally agreed model for second-language writing pedagogy or scholarship; as Raimes (1991) discusses, in many areas scholarship in L2 composition over the past 25 years has not taken us very far. However, Raimes suggests that there are some "emerging traditions" in the teaching of second-language writing: recognition of the complexities of composing, of student diversity; of learners' processes; of the politics of pedagogy; and of the value of practice relative to theory. While not all scholars and practitioners in L2 writing would agree with Raimes' position on these five "emerging traditions," they would probably agree with her that these are all issues to be taken into consideration in the development of new forms of assessment of ESL writing.

Relating Theories in Composition to the COE Model of Communicative Competence

From the brief overview we have provided of the present theoretical orientations of L2 composition studies, it can perhaps be seen that the TOEFL Committee of Examiners' (COE) model takes a quite different starting point. This model begins from the notion of communicative competence and proposes to identify the academic domain of language use as consisting of an interlocking dynamic that can be broken down into eight factors: settings, text types, tasks, procedural competence, linguistic competence, discourse competence, sociolinguistic competence, and functions. Though the model divides the universe of writing into these eight components, we have found that because the terminology used in discussing both communicative competence and the COE components is not found in the field of writing itself, these concepts do not help to illuminate the issues as clearly as they appear to in the oral sphere. We have therefore looked for points at which we could connect the concepts in the COE model with the organizing conventions that have been proposed in rhetoric/composition studies.

The notion of "purpose" is one commonly used in L1 composition, having its genesis in the work of Britton et al. (1975), which divides all acts of writing according to the purpose of producing a text, for example writing to learn, writing to display, writing to persuade, and so on. "Purpose" is also illustrated in the COE model in the category labeled "functions," which lists a variety of uses of writing, together with the COE category "tasks," which more narrowly lists things that one might accomplish by writing. To apply the notion of "purpose" to the TOEFL 2000 project, one might attempt to identify all of the potential purposes for which writing is used in the academic environment and then designate which purposes would best lend themselves to being tested as a way of sorting students by proficiency levels. But as Ackerman (1993) points out, the work of Britton et al. (1975) and the British Government report *A Language for Life* (1975) placed the writer's abilities and needs *above* the subject matter and language demands of textual expectations, seeding the Writing-Across-the-Curriculum movement in the U.S. with expressivism rather than genre. In composition studies, "purpose" belongs with audience as we teach writers to think about those who will read their text as part of their purpose; a text is for people rather than for things. This purpose and audience nexus, so fundamental in the teaching of academic writing, may be implicit in the COE model but is not explicit; indeed, co-occurring variables are generally hard to account for within the model. Further, as we think about purpose more broadly in the academic contexts into which TOEFL examinees will enter, we will see that learning purposes calling for writing rarely exist wholly separated from purposes involving wider language functions. This is recognized both in the Britton et al. original research (1975) and in the national report *A Language for Life* (1975) that resulted from this research and greatly influenced British education in the 1970s. From this perspective then, the possibility of integrated testing of writing with the other language skills must be seriously considered (see the following discussion).

Another approach that might be taken in designing a writing test component for the TOEFL 2000 project would be based on text analysis. One could generate a model of writing by identifying the dominant features that could be used to describe either idealized or real examples of different text types. This is the principle behind the traditional division into the four rhetorical modes of narrative, descriptive, expository, and argumentative prose, or to a lesser extent, the traditional labels for the English prose patterns of comparison/contrast, classification, cause and effect, and so on. Both of these approaches are now viewed as flawed (Connors, 1981). "Text analysis" overlaps considerably with the COE category of "text types," but also includes aspects of the COE lists found under "procedural competence," "discourse competence,"

and "functions." To apply the notion of "text analysis" to the TOEFL 2000 project, one might determine that students need to demonstrate their proficiency in English by being able to produce a series of texts that adhere to conventions typically associated with a variety of different text types. Each student sample could be read and evaluated for presence or absence of the critical features a textual analysis might identify in a well-formed English text responding to the same task.

Both of the preceding examples of models fail to take into account the entire context in which academic writing is produced and the entirety of how it is evaluated in authentic academic settings. They run the risk of resulting in inauthentic assessments that privilege the perspective of English teachers over that of subject faculty who are most likely to encounter these writers in academic coursework once they have been admitted to U.S. colleges (see Hamp-Lyons, 1991a; Johns, 1990). Yet, other models that play a critical role in the field of composition studies may seem unhelpful because they are not so much models of writing as a product as they are models of writing as a process. The influential Flower-Hayes cognitive process model (Flower & Hayes, 1981) divides the writer's world into three parts: the task environment, the writer's long-term memory, and the writing process itself. While this model is rich in the explanatory sense, and the foregrounding of "task environment" is appealing to writing assessment researchers who know how large a part writing tasks and prompts play in explaining the results of writing assessments, the prominence it gives to long-term memory and to the writing process is problematic for the design of academic writing assessments. Equally problematic is a model that privileges expressive writing, for it leaves out an account for contexts within which writers make strategic decisions and construct texts. What is needed, we believe, is a model that brings together all the elements in the theories and models we have outlined above.

Drawing on the work of Berlin (1982, 1987), Johns (1990) offers a helpful interpretation of how theoretical models interact with classroom decisions. And perhaps by extrapolation, we can apply her methodology to a model for ESL writing assessment. Her succinct summary of the heart of Berlin's work is as follows:

> [Berlin] suggests that all complete rhetorical theories and, by extension, all approaches to teaching composition must consider the following: (1) the writer (or "knower"), (2) the audience (or reader), (3) reality and truth, and (4) the sources of language in written text (Johns, 1990, p. 24).

She then goes on to detail how these four components can provide the organizing principles for a presentation of the three different approaches to writing in Berlin's model (discussed earlier), namely process approaches, interactive views, and social constructionist views. Each of these approaches provides a different vision of the four components of Berlin's model, resulting in privileging one or another component in the composition curriculum that derives from a particular philosophy. If we apply Berlin's model to writing assessment using Johns' framework, we can see that testing instruments for ESL students will differ greatly depending on what is to be tested. A test that seeks to discover whether students are able to use language accurately and to structure texts in rhetorically suitable ways might be a very different instrument from one whose primary goal would be to generate samples for evaluating students' ability to

demonstrate their membership in a particular discourse community (writing like a biologist or writing like a historian, for example).

If we wish to assess the discourse (rather than the communicative) competence of TOEFL 2000 candidates — that is, if we wish to assess global writing proficiency — all of Berlin's categories must be factored into the test design. We use the term "discourse competence" because we believe this term captures the notion of writing as an act that takes place within a context, that accomplishes a particular purpose, and that is appropriately shaped for its intended audience. We cannot, we believe, consider a writing test to measure a skill apart from its reason for being. That would be like trying to measure reading without comprehension (as, for example, in pure decoding) or speaking without understanding (as, for example, rote repetition of sounds heard).

Five sets of questions were constantly in our minds as we shaped the preceding discussion, and must inform our own thinking and that of our readers in the following discussion of issues in designing an appropriate writing instrument:

1. What universe of writing skills is needed to succeed in an academic context? To what degree are these skills socially constructed, and how much are they the results of expressive and cognitive processes that are relatively decentered from specific academic tasks or expectations?

2. Which of these writing skills are needed by undergraduates and graduates, the two main academic populations served by the test? How can we elicit samples of those writing skills in a way that is true to the nature of "real" writing and to "real" writing in academia, and that is appropriate for administration to large numbers of test takers?

3. Are the performance expectations for those skills the same for the two populations? In examining performance expectations for the purpose of determining scoring procedures, whose expectations should be privileged: those of English instructors, of faculty teaching community college and lower-level courses at four-year schools, or of faculty teaching their disciplines to subject majors and/or graduate students?

4. How can we judge the writing skills demonstrated in a way that is organizationally feasible? How can we prepare readers to judge writings that are appropriate to the writers' purposes and their constructed audience (as opposed to the actual audience, the test essay rater)?

5. If we succeed in eliciting and evaluating samples that are congruent with the expectations of the academy, how are we to avoid disadvantaging test takers who do not have experience in the expectations of the academy?

We will not claim to have answered these questions, only to have raised and explored them as far as the current status of the field takes us, for as Johns (1990) reminds us:

> . . . because world views among theorists, researchers and teachers in both the first language and ESL differ, . . . no single, comprehensive theory of ESL composition can be developed on which all can agree (p. 33).

Exploring the Nature of Academic Writing

Because of the purpose for which, we are told, the TOEFL 2000 is to be used, and because of our own theoretical location relatively toward the social constructionist end of the spectrum among the models discussed above, we begin our discussion of the specifics of test design issues with a consideration of the nature of academic writing.

Contextualizing the Academic Writing Situation

The TOEFL 2000 working draft (April 1993) describes TOEFL 2000 as "a measure of communicative language proficiency in English (focusing) on academic language and the language of university life." The Committee of Examiners began work on a model of language use that could support test design. In this model, "academic context" has been subtitled "observable situation" and is shown as composed of situation and production output. Situation is itself made up of task, text, and setting. Missing from the features of situation listed in the COE model is "expectations," which both studies of writing assessments (Hamp-Lyons, 1991a; Johns, 1991b) and case studies of individual students within the academy (Herrington, 1986; McCarthy 1987) have shown play a very large part in determining the value placed on student/test-taker texts (what is referred to as "production output" in the COE model). The listing of possible settings, text types, and tasks includes the range of such situational realizations encountered in the academy. The problem here is that no individual test taker will encounter or have previously encountered the whole range shown, and, of course, some or most of them will not have encountered any (particularly candidates for undergraduate admissions).

There are three ways to resolve the issues of situational responsiveness: (1) develop several discipline-focused tests, with realistic situational variants used for each, as was attempted with the English Language Testing Service (ELTS)[1]; (2) create very broad distinctions that will hold across a wide set of disciplines, such as the humanities/sciences distinction used by the International English Language Testing Service (IELTS), the successor of ELTS, designed at Lancaster University and implemented in 1989, and the Test of English for Educational Purposes (TEEP), designed in the early 1980s by Cyril Weir for the Associated Examining Board, Aldershot, England; or (3) attempt to extract the most commonly occurring situational elements and design a single test structure to test those common situational aspects. The current COE model sits awkwardly among these, providing as it does a list of all possible settings, text types, and so on, without an organizing frame that would enable test developers to know which of these realizations applies to each of the possible separable test populations. To take just one example, listing "lab reports" and "book reviews," is not helpful unless the academic contexts within which these are common authentic tasks are also stated. As some assignments submitted to the Hale et al. study (1996) illustrate, even book reviews can be solicited from students with widely ranging sets of expectations by professors of different discipline.

This is, of course, a content validity issue. Anastasi (1982) sets up the following requirements for establishing the content validity of a test.

1. The behavior domain to be tested must be systematically analyzed to make certain that all major aspects are covered by the test items and that they are in the correct proportions.

2. The domain under consideration should be fully described in advance, rather than being defined after the test has been prepared.

3. Content validity depends on the relevance of the individual's test responses to the behavior area under consideration, rather than on the apparent relevance of item content.

The COE model of language is not empirically derived and thus cannot help in this inquiry. However, the Hale et al. (1996) study provides a data base from which it might be possible to seriously consider developing a writing component for TOEFL 2000 that will provide reasonable content, whichever option of specificity or generality is chosen, as far as requirements (1) and (2) are concerned. Aside from the data collected for the Hale et al. study, the most comprehensive study of academic language use to date was carried out by Cyril Weir (1984) for his dissertation and involved a more broadly based data collection from academic contexts than the focus on formal assignments chosen by the Hale et al. study. Weir carried out an in-depth content validity study of writing in British universities, which included questionnaires to faculty, questionnaires to students, detailed in-class observations, and the development of a framework of communicative test events that was then applied to the development of the TEEP, based on the questionnaire responses and classroom observations. Although Weir's study was carried out in Britain, and there are a number of significant differences among the educational systems and expectations in the United States, Canada, and Great Britain, it could still provide a starting point for research studies at North American colleges and universities. However, to complete the picture of content validity and to respond to requirement (3), studies of test takers as they respond to items at varying levels of specificity and generality in domains closer to and further from the individual's own field(s) of expertise are needed. While very little of such research has been done, studies by Cohen (1984), Cohen and Cavalcanti (1990), Connor and Carrell (1993), Hamp-Lyons (in press-a), come to mind.

Writing Tasks and Genre Issues

Without an understanding of the nature of academic writing tasks, it would be impossible to construct a test that would meet even the minimum test of face validity, namely requiring test takers to produce one or more writing samples that are similar in nature to the writing they are asked to produce in courses they enroll in. The current Hale et al. study (1996), which is a survey of academic writing prompts, should prove helpful in this regard. The study suggests a classification scheme that allows us to identify the exact nature of the writing tasks being demanded of students in eight disciplines with heavy ESL student enrollment. This classification scheme, which has now been applied to several hundred writing assignments collected from the eight campuses targeted for the study, has five categories:

1. locus of writing, i.e., in class or out of class
2. length of product to be written
3. genre of product
4. cognitive demands of task
5. rhetorical properties of product, subdivided into mode of discourse and specification of rhetorical patterns

It seems that three of these factors are particularly critical to our discussion: genre (to be discussed here), cognitive demands, and rhetorical specification (to be discussed later as they relate to difficulty level in prompts). Missing from the data collected was information on whether the writing done in test situations was "impromptu," that is, on a topic that not only did the writer not know in advance but that was also not part of some immediately prior and related educational experience from which the writer could reasonably be expected to draw. Presumably, students are never given "impromptu" writing tasks in subject course assignments and tests, but because such tasks *are* traditionally given on English proficiency essay tests and will doubtless be an option considered for TOEFL 2000 writing, it will be important to evaluate the level of reality or unreality they represent.

In the Hale et al. (1996) study, the genres by which assignments were classified in the coding of raw material were derived from the actual assignments and test prompts themselves and ranged from the rather generic "essay" to the very specific and narrowly defined "documented computer program." It seems possible to try to identify all of the possible genre types found in student writing (a sample listing is found under "text types" in the COE domain scheme), but we must then consider which of that universe of genres could possibly be demanded on an exam. Clearly, to ask someone to produce a field-specific text, such as a documented computer program or a case analysis in a business framework, is to presuppose that the test taker has already worked with that genre (and thus has formal schemata to draw on), has content knowledge (content schemata) to inform his or her writing, and can produce the requested text type within a time framework appropriate for a testing situation. Even if all of these conditions were met (and this is highly questionable), this would result in a highly fragmented test, with a very large number of choices for test takers deciding which genre they wanted to be examined in. Such problems, in addition to the extreme test development demands, were behind the British Council/University of Cambridge Local Examinations Syndicate's (UCLES') decision to reduce the number of models in their new IELTS test from the six to two used in the original ELTS writing test. We shall return to the problems of such a model shortly.

Some other genre types that were found in some assignments submitted to the Hale et al. (1996) study also seem inappropriate for mass testing since they would have resulted in products for which it would be very difficult to develop scoring guidelines that distinguish among the proficiency levels of the test takers. Among the assignments collected, for example, were free writing samples, journal entries, and e-mail messages. Other genre types that are quite commonly found in course requirements in a variety of disciplines are inappropriate for testing situations because the preparation needed to produce them requires too much time. These would include, for example, independent library research into multiple sources for a paper requiring synthesis and interpretation of large quantities of material, or reading a book-length text in anticipation of preparing a critique. This does, however, leave genres other than the "bare" essay that we have seen in real academic writing that could be adapted to the testing situation, such as summaries of written texts, comparisons of short texts, letters, and appeals/petitions, all of which are also identified on the COE scheme under "text types." In any case, it would appear from a preliminary look at the assignments submitted for the Hale, et al. (1996) study that it is difficult to separate the issue of genre from the issue of the difference in undergraduate- and graduate-level writing tasks and requirements as found in institutions of higher education.

What complicates the picture, apart from distinctions at the graduate versus undergraduate level, is the uncertain nature of disciplinary boundaries themselves. For example, it is becoming increasingly common

to talk of "disciplinary cultures" (e.g., Becher, 1987) and to question a view of academics as members of a single profession, tending rather to viewing "the academy" as a large number of different professions. This view was shared by those who designed and introduced the British Council's ELTS and led to the test's multi-option structure. There is, however, no set categorization of academic disciplines into rational classes by, for example, research modes, language genres, or other clear differences. Currently, then, a problem with attempts at disciplinary classifications is their lack of generalizability, and in this context a more general model of academic language is frequently used to underpin test development. Alderson (1981) points out that a priori "a specific test is impossible" (p. 123): it is not possible to construct a test for every describable group of potential test takers, and we do not believe TOEFL 2000 can offer an infinite variety of possibilities. Furthermore, Becher points out that in describing the content of academic disciplines there are two opposing tendencies, one tends to reduce the arena of investigation to a manageable size and the other insists on the recognition of important distinctions even within a single discipline.

Process Versus Product

Many current writing assessment tests, such as TWE (Test of Written English), WAT (Writing Assessment Test), used by City University of New York, and the writing portion of both the University of California at Los Angeles' ESLPE (English as a Second Language Placement Examination) and the University of Michigan's MELAB (Michigan English Language Assessment Battery), appear to be tests that judge only written products. Yet it is inaccurate to describe them as tests of product, since the only way the test taker can present a product to be judged is through the (successful or unsuccessful) writing processes the test taker applies to the product readers will evaluate. Interestingly, Wolcott (1987) has pointed out the contradictory nature of two developments in the field of writing in the last decade: the emphasis on writing as not only a vehicle of expressing meaning but also of discovering meaning; and the simultaneous expansion of writing assessments, which typically focus on written products. We are not talking here about the writing processes commonly taught in process writing classrooms, with their range of prewriting techniques, drafting, and revising, and collaborative processes such as peer critiquing and conferencing, for writing can and does go on, successfully in some cases, without them. Rather, we are talking about unavoidable processes that are inherent to all text-making. Awareness of the complexity of writers' processes, it seems, has been a key force in winning the battle for direct writing assessments over multiple-choice testing. Now, the same awareness, coupled with the impetus toward acknowledging the needs of diverse writers, is leading us toward portfolio assessment (Hamp-Lyons, in press-b).

A distinction often made about the genesis of text is that of knowledge-telling versus knowledge-making. Since knowledge-telling requires access to the material or information from which knowledge is to be drawn and reported, a writing test must avoid all knowledge-telling tasks unless the raw material from which the knowledge can be obtained is provided within the test. Thus, except in writing tests that are integrated with other language skills and the information is provided through the medium of one or more other language skills (reading and/or listening), almost all responses to writing prompts are likely to be knowledge creation. This can be seen in the type of "bare" prompts (i.e., those stimuli without reading material that supplies self-contained context in relatively few words) that can often more or less

be reduced to a formula that asks the writer to "Give your own reasons for . . . [and] support your answer with examples and details." Tasks like this put heavy emphasis on the processing component as described in the Grabe revisions of the COE model, requiring on-line processing of the prompt (what Grabe calls "the sensory input") and integration of the proposition put forward by the prompt, first through language comprehension processes and then through critical faculties through the schematic network in order to access conceptual and content schemata appropriate for application in response to the prompt. After this, the test taker still has to consider issues of context, audience, and genre, choosing from the accessed possibilities those that she or he feels will be acceptable to the (unknown) tester and that are within the realm of the test taker's English language competence. Often too, constraints of the test situation require the test taker to select material and strategies that construct a task that can be done in what is usually the rather limited time available (remaining) and can be organized adequately in a first (and only) draft. (These issues will be revisited later in the full discussion of prompt development.)

Test takers who do well on the highly time-constrained tests are likely to be not only those with a strong linguistic competence but also those who have had exposure to writing in timed test situations, and who already possess a well-formed "model text" and "mental model" of the test (to borrow Grabe's terms). These demands are heavy for a 30-minute or 60-minute time period, and the fact that so many writers are able to write successfully under these conditions suggests that compensatory strategies are at work. For the TOEFL 2000 we must work to design test parameters that call on writers to apply their discourse, sociolinguistic, and metacognitive abilities as well as their linguistic and rehearsed genre competence in an entire package. We must design a test that acknowledges the writing processes all written products depend on. If we do not, the test will only give information about some competencies, and those competencies may not be the ones about which we most need to know in the context of TOEFL 2000 goals. Furthermore, for the sake of the sociopolitical context in which TOEFL 2000 will exist [and the foreshadowing of this context can be seen in the papers by Greenberg (1986) and Raimes (1990)] as well as for the best interests of the test takers, the design of TOEFL 2000 must be responsive to concerns about writing assessment already being raised by others, at ETS as well as elsewhere. Camp (1993) explains a basic assessment struggle as follows:

> What we are experiencing . . . is a mismatch between the complexities of the conceptual framework for writing that we find in current research and practice and the simpler construct implied by traditional approaches to writing assessment, including the writing sample. Very likely we are also seeing the signs of a growing incompatibility between our views of writing and the constraints necessary to satisfy the requirements of traditional psychometrics (p. 52).

Models of Academic Writers

Empirical studies may allow for the establishment of a model of academic writing proficiency that specifies the key characteristics of successful "apprentice writers" that can be shown to be common across disciplines. Herrington (1986), looking only at the experiences of undergraduates and only within an L1 context, suggested some such characteristics. Successful student writers perceived that in their writing they needed to create an issue for themselves and work to resolve it, first for themselves and then to convince their professors; they saw themselves as an audience in the sense that they use their writing to explore an

issue and shape their responses and convince themselves that they had resolved it to their own satisfaction; and they were able to interpret and act on the information they received from their professors in getting closer to an understanding of the disciplinary culture, even though professors often left tacit much that was central to the value system of the discipline. In a more recent paper, Johns (1991b) has argued that Tran, a student she followed in a case study, was unable to pass his English competency exam despite getting high grades in biology and chemistry courses, in part because the tasks on the competency exam were so unlike the written tasks he was used to in his science courses. Hamp-Lyons' experience at the University of Michigan, however, showed that many of the written tasks collected from her students (usually of the long essay type) allowed great freedom to select a topic and nominate a way of working with it: "task representation" was to a considerable extent in the hands of the writers. Similarly, at the University of Edinburgh, Hamp-Lyons (1986) carried out a small-scale study of the design parameters faculty have in mind when they prepare tests of writing in their own discipline, and of the criteria they use when scoring the writing produced by the graduate writers on their courses. The design factor these 24 faculty most frequently cited was "avoidance of ambiguity." The next most frequent design factor, again variously worded, had to do with finding out what students can do with knowledge. Here, faculty talked about allowing students to show the ability to apply knowledge to problems in the discipline; about getting students to show discrimination; and about designing tasks that required students to discuss or argue, not just display, knowledge (see Horowitz, 1986, for a categorization scheme that uses several of these terms). These British faculty were clearly concerned about knowledge-making and not knowledge-telling. Data in the Hale et al. study (1996) are beginning to suggest that graduate-level students are often asked to apply their own interpretative frameworks to knowledge acquired in the service of generating original concepts. To participate in the creating of knowledge is widely considered empowering. Tran (Johns, 1991b), on the other hand, wanted tasks that permitted him to display, rather than generate, knowledge.

Because the distinction between knowledge-telling and knowledge-creation is one that is often made along undergraduate-graduate lines, it is related to another issue that must be considered for TOEFL 2000 writing — whether the same test can be used for both undergraduate and graduate students. The Bridgeman and Carlson (1983) study found considerable differences between undergraduate and graduate tasks and expectations. The current Hale et al. (1996) study should provide more information about such differences. Expectations in key areas such as breadth and depth of reading, ability to generate an original argument or thesis, the amount of support required for an argument, and so on, all appear anecdotally to differ markedly, and now we are getting some empirical evidence that will support or refute the validity of this view. Furthermore, we should not overlook the fact that when undergraduates taking the TOEFL 2000 achieve their (likely) objective of acceptance to a particular institution, they are most likely going to have to complete one or more required writing courses as part of their undergraduate curriculum. They should not be held accountable for demonstrating the type of proficiency that they are expected to be taught and to learn *after* the exam is behind them. Conversely, graduate students taking the TOEFL 2000 for admissions purposes will not necessarily find institutions that either require or even have available writing courses designed to increase their proficiency; rather they are expected to have a threshold level of writing competence upon entrance to the program of their choice.

In this vein, Swales and Feak (1994) argue that the writing needs of graduate students are distinctly different from those of undergraduate students, and at the University of Michigan, under Swales'

influence, undergraduate and graduate students are tested on different instruments, the Academic English Evaluation for undergraduates and the Graduate Test of Academic Skills in English for graduates. In the United States, perhaps more than in many other countries, undergraduate education is general education; students typically need to become "initiated," in Bartholomae's terms (1985), to several disciplinary discourse communities even in a single term. At the graduate level a student's studies are far more focused, the boundaries of what is known and what is neither known nor expected to be known are more clearly delineated. On the surface, then, developing a test that will be fair to all graduate students and yet have some semblance of academic specificity will be more problematic than developing a similar test intended only for undergraduates. Further investigations should add to the small body of empirical data currently available on the actual writing experiences of second-language undergraduates and graduates within the academy before a principled decision can be made about the appropriacy of a single measure of writing for both undergraduates and graduates (see also Purves, 1992).

Assessment Variations

Assessment for Communicative Competence

As Savignon (1991) notes in a review of communicative language teaching in ESL, the term "communicative competence" does not lend itself to "simple reduction" (p. 263) but rather remains "a robust and challenging concept for teachers, researchers, and program developers alike" (p. 263). The challenge for the assessment community, in keeping pace with changes in classroom approaches, is to find ways to create tests that allow students not merely to demonstrate their mastery over fixed pieces of the language abstracted from their possible communicative function but rather to demonstrate their level of proficiency as language users in action. This is clearly the challenge to which TOEFL 2000 must address itself.

The COE has provided a hypothesized model of communicative language use in academic contexts that includes a list of text types found therein. A fully developed model will include both the genre features of each text type and a set of contextual features that describe the conditions under which each text type is created. This would go a long way toward helping to concretize the concept of "discourse competence." Lab reports, for example, are first written up as notes during experimental observations in a lab, and then usually written in appropriate format and style later. Book reviews too begin as notes made during an observational process — in this case, the reading and interpreting process — and are later revised, expanded, and integrated with a theme, position, or other type of narrative thread into a conventional genre. Appropriate text/task types for the writing component of TOEFL 2000 can be chosen from a fully developed model that can identify not only a genre but also the framework in which each genre becomes actualized.

When a test asks students to demonstrate their membership in the community of fluent writers of English, it should certainly hold those students accountable not merely for the formal properties of the text they produce in terms of rhetoric and syntax, but also for the entire communicative act that is embodied in their writing sample. Janopoulos (1993) speaks to these issues when he claims that ESL writing teachers "*place a premium* on evaluating a student's efforts in terms of how effectively that student is able to communicate" (p. 307, emphasis added). However, when students are not held accountable for the truth value of what they write or for the relevancy of the content they use to address a specific task at hand, something is lacking in discourse (and communicative) competence. If TOEFL 2000 is to be seen as a test of communicative language, then this concern must be addressed in the prompt specifications, in the scoring criteria and rater training, and in the publicly available documentation about the writing test, so that a premium is not placed on any isolated component of the text(s) produced by test takers. At the same time, the questions of how well test takers will be able to construct for themselves a valid sense of purpose as this is defined for them by the test specifications and ETS-distributed documentation and training materials must be addressed. Many TOEFL 2000 test takers will have no experience of either North America or North American colleges and universities. This makes not only the purpose component but also the audience component very difficult for them (Hamp-Lyons, in press-a).

A Snapshot Approach

One might refer to the single-sample text collected under severely limited time constraints on a topic the test taker has not prepared to write about as a "snapshot" approach to testing. Such a method of eliciting writing samples cuts writers off from much that is part of their writerly skills, if they have them. A potential consequence of this is the compression of apparent writing abilities into a smaller score range than might occur if writers were able to find their own level by writing on topics they felt comfortable with and in an amount of time that (while still constraining) allows room for those with a good deal to say and/or extensive composing skills to show what they are capable of, while reinforcing the impression created by weaker writers that they have little to say and/or are constrained by limited language command from saying it. Other consequences are loss of construct validity (the behavior being measured is in important ways unlike the normal behavior of people performing the skill, especially for the more skilled writers), loss of face validity, and as suggested in the previous section, a resulting inability to report test takers' proficiency in writing on meaningful tasks. In addition, much research (e.g., Ruth & Murphy, 1988) has indicated variability in test takers' measured ability resulting from differences across prompts and prompt types.

Any writing component that contains a single sample of writing will be prey to many of the same problems that have been identified by the TWE test and other testing programs that elicit a single sample: What type of writing? How narrowly should the prompt characteristics be constrained? How to extrapolate from such limited data to test takers' performances on authentic academic writing tasks?

A Growth/Multiple Competencies Approach

Currently, in the field of education generally, educational assessment (especially as represented by the new journal *Educational Assessment*) and composition pedagogy in particular, the use of portfolios is attracting great attention (e.g., Belanoff & Dickson, 1991; Yancey, 1992). Portfolios have been heralded as a favorable assessment approach to the writing of ESL students (Brookes, Markstein, Price, & Withrow, 1992; Valdez, 1991) because of the benefits of extra writing time, access to support services such as writing centers, and for other reasons described in detail by Hamp-Lyons (in press-b). To date, however, no empirical research has indicated that ESL writers gain an advantage from assessment by portfolio rather than the traditional writing assessment, and Hamp-Lyons (1993) reviewed a number of studies that raise theoretical or empirical cautions about the value of portfolio assessment with ESL writers. While the authors are in favor of portfolios in general, it is clear that introducing a portfolio writing assessment for TOEFL 2000 would be as problematic as (and no doubt more costly than) introducing a computer-delivered multiple-choice component for TOEFL 2000. The authors tend to favor a more modest approach, perhaps incorporating some elements typically found in portfolio assessment; the minimum would be a multi-sample exam.

Assessing Writing as Academic Literacy

Great concern is being expressed, not only in the media but in educational circles as well, that many young people leave high school and enter college without having achieved a level of literacy that will make them fully functional in academic contexts. Many colleges are finding that, to the long standing

freshman writing and basic writing courses, they are having to add courses in academic reading. High school graduates rarely read for pleasure or for self-initiated learning; they read slowly and/or without the background knowledge that makes appreciation of metaphor, allusion, and exophoric reference possible. They are handicapped in their writing courses because these courses so often work from readers or use outside reading as stimulus for writing assignments (see Carson & Leki, 1993, for multiple discussions of this issue within an ESL framework). Concern over these issues has made "academic literacy" a recognized term and an official goal on many college campuses. Writing, we know, is not (at least not in its academic uses) a stand-alone skill but part of the whole process of text response and creation; when students use both reading and writing in crucial ways they can become part of the academic conversation — they signal their response to academic ideas and invite others to respond to their ideas in turn.

The richness of such a view of academic literacy suggests that in a TOEFL test for the new century an attempt should be made to design at least some integration between reading and writing skills. This could be done without prejudicing a fair measure of pure writing skill and reading skill by, for example, using one reading text as content input to a writing prompt, or by following a "take a position" essay with a reading that takes one specific position on the same issue. Research would be needed to estimate the effect on test takers' scores from the doubling of skills. Weaker writers might be disadvantaged by prompts that presuppose comprehension of a reading, or weaker writers could be helped by a reading that would provide them with some information for their writing, some vocabulary to mine, some genre conventions to model a response upon, and so on. In a multiple-item writing test, the effect of a text-linked writing task could be discovered.

Meanwhile, in first-language composition research, a growing body of research has been exploring the ways in which student writers use source texts in the production of papers (see, for example, Flower, Stein, Ackerman, Kantz, McCormick, & Peck, 1990; Kantz, 1990; McGinley, 1992; Spivey, 1990). More such studies with ESL students might be conducted, both specifically along the lines of the work done by Campbell (1990) on how students use material newly presented to them or by Sarig (1993) on composing a study-summary, and more generally, by investigating issues of academic literacy in the ESL context. Collignon (1993), for example, argues that at the community college level, ESL students need "literacies" (Gee, 1989) instead of the narrower "literacy," because they seek to discover culturally appropriate ways of using language(s) as well as to manipulate discourse modes within a language.

Assessing Writing Within Wider Academic Competencies

While reading and writing are often viewed as key language skills for success in the academy, speaking and listening are also vital. No student will succeed in a lecture context without formal listening abilities; any student whose spoken command prohibits him from visiting a professor's office and asking questions about assignments will be at an extra disadvantage in the class. There is, then, *prima facie* validity to the notion of an integration of all skills within a performance-based TOEFL 2000. From the point of view of authentic testing of writing, this would make a great deal of sense, since academic classrooms and writing classrooms in academic contexts are multi-skill environments, where talk is used as the essential underpinning of all the work of writing. Indeed, verbal negotiation and the search for shared meaning characterize the process approach to writing instruction. The University of Michigan's Tests and Certification Division in the English Language Institute has developed and implemented a multi-skill

test for graduate students who are nonnative speakers of English, the Graduate Test of Academic Skills in English (GTASE). Some research into the feasibility of a four-skills test for TOEFL 2000 would be worthwhile. However, the same issues of skill confounding referred to in the immediately preceding section would need to be investigated in this context.

Additionally, it is worth keeping in mind that there are academic skills that cannot be satisfactorily defined as only one skill or another; note taking and note making come most immediately to mind. Note-taking involves listening and writing (or, at least, some traits of writing, such as organization), and testing these separately would not add up to a note taking test. In the same way, note making involves both reading and writing. A content validation study is clearly called for in this area and it would need to go beyond the listing of possibilities or the gathering of paper documentation on academic learning contexts to class and individual observations, such as those carried out by McKenna (1987) and Chan (1990).

Writers as Test Takers

We understand a great deal less about our test takers from countries around the world than we need to. This is the great underresearched aspect of language testing. The field of composition has writing in contrastive rhetoric to draw on, but the status of contrastive rhetoric research is in some question (see Leki, 1991, for a recent review). The remarkable paper by Shen (1989) has revealed for us how conflicted a second-language writer can be in the North American academic context — but Shen is clearly a successful, indeed an exceptional, student. Reid's studies (1990, 1992) of the performance of four different language groups on the prototype essays for the TWE test also illustrate with some detail observable differences in the performances by students from different language backgrounds.

Differences among the test takers are both natural and desirable, since the test has a discriminant function, but these differences must be due to real differences in English-language writing ability by the test takers and not due to either obvious or subtle bias in the test. Bias most obviously can occur in the type and wording of prompts, but bias may also be present in the scoring guide or in the raters. Both of these issues are discussed in more detail in later sections. In the research and development agenda for TOEFL 2000, it is noticeable that, while many constituencies are courted, the test takers themselves, the ultimate stakeholders in this testing context, are almost excluded from advisory meetings and from any role in the decision-making process itself. Yet within the field of education there is a groundswell of interest in collaborative assessment and self-assessment, and both of these deserve our attention as we work toward TOEFL 2000. We could learn, for example, from Cohen's work on test takers' strategies and on their responses to test formats (Cohen, 1984; Cohen & Cavalcanti, 1990).

There are no established strategies or research methods for learning about the writers who take our tests, and indeed all discussion of writers as people is usually completely missing from any evaluation of a writing test. Tests of English as a second language generally analyze easily obtained demographic data such as native language and country of origin, perhaps language proficiency on a more general measure, then look at writing test performance against these variables. This lets us understand something about how writers as groups are affected by test variables, but it precludes understanding of the many individual factors related to background, experience, and personality.

Each writer brings the whole of himself or herself to the essay test. Each writer is a complex of experience, knowledge, ideas, emotions, and opinions, and all of these things accompany the writer to the essay test. In interpreting and responding to the topic of an essay test, each writer must create a fit between their world and the world of the essay test topic, and each must find a way of making sense of the task before she or he can respond to it. In the context of classroom assignments, Ackerman (1990) notes that "in many cases the assignment given by an instructor and the assignment taken by a student are not a reciprocal fit" (p. 96), further pointing out that "giving and responding to an assignment is an act of negotiation" (p. 96). There is every reason to believe the same is at least partially true on "assignments" that form the prompts on exams requiring essays. When the topic/task is a very wide one, it is, as Labov (1969) said, "absurd to believe that an identical 'stimulus' is obtained by asking everyone the same 'question'" (p. 108). Murphy and Ruth (1993) review some specific examples to illustrate this point.

There is some research, with first-language high school students (O'Donnell, 1968) and with ESL students (Hamp-Lyons & Mathias, 1994), that suggests weaker writers are less successful at making

good choices about writing prompts when choice is available, and this is an issue that requires further understanding. When the task is a very narrow one, writers whose personal histories have the closest "fit" to the expectations of the task will find it easiest to interpret. Each writer needs both guidance on what is important about this writing task and what qualities will be valued, and some room in which to maneuver in taking the task and topic and creating an original, personal response. Pollitt, Hutchinson, Entwhistle, and De Luca (1985) refer to this as "outcome space" (p. 7). The more constrained the outcome space available to the writer, the less room there is for the writer to reveal some unique quality of herself or himself, some unexpected strength or weakness; and the more room there is for the writer to fail to write within the boundaries of acceptability. Yet the more freedom there is, the less support the writer can get from the prompt itself, the more the writer is thrown onto her or his own resources, and the more seriously she or he is disadvantaged if those resources are limited in this area. Regrettably, we have not yet achieved a rule of thumb for identifying the degree of freedom and constraint that allows writers to show their best selves. However, Brossell (1983) found that topics with a moderate level of rhetorical specification (specification of purpose, audience, voice, and content) yielded higher mean scores than essays with either a high level of specification or a low level of specification. There is a logical appeal about his finding that is worth giving consideration. Golub-Smith, Reese, and Steinhaus (1993) further explore differential test taker performance based on the notion of implicitly versus explicitly stated tasks as exemplified in a small sample of TWE prompts.

Test Variables

In this section, we will review some issues raised in connection with the various aspects of a writing test. These aspects have been well-rehearsed in the field of writing assessment and are summarized by Hamp-Lyons (1990) in the context of ESL testing.

Prompt Development

In large-scale testing, it is critical for the validity of the test that the test takers be given a test that is equivalent from one administration to the next. In multiple-choice tests, statistical procedures exist for assigning a value to each item in the test to represent, more or less, its difficulty level (based, of course, on field testing the items). Any given exam consisting of many different items can thus be constructed from items that have a range of difficulty levels, and that is one way to assure that an exam administered to test takers in January can be deemed similar to an exam administered to test takers in all the other months of the year. In a writing test that has a single-item prompt to which all test takers must respond, this validity can only be derived from providing prompt after prompt, in which each prompt is somehow perceived to be equivalent in its accessibility and difficulty levels both to previous and future prompts. Even the most sophisticated statistical procedures now available require some crossover — at least some test takers must do multiple items — for equivalence to be estimated. This is clearly a very important factor in shaping the written component of TOEFL 2000.

But what do we mean when we ask if one writing prompt is similar in difficulty level to another writing prompt? One could, for example, specify a range of prompt features that can be controlled (e.g., linguistic complexity, genre demanded, rhetorical specification, audience, subject matter, and so forth) in order to develop a set or sets of presumably parallel prompts, which would then have to be pre-tested and statistically equated (and rejected if they failed the equating procedure). Prompt equating at pre-test would seem to be a basic necessity for any writing test, but it cannot be done properly unless (a) very large pre-test samples are collected, all of which would be read and evaluated, or (b) a multiple-item writing test including pre-test items, as presently happens with the TOEFL test, were to be used to create a bank[2] of items. Comparison of score patterns of pre-test items with those test takers' scores on actual items would enable go/no-go decisions to be made. The third alternative, if a multiple-item test is used, would be to apply FACETS analysis operationally, as is discussed later, and as is done on the Australian ESL test "**access:**." Further, writing prompts should also be matched or equated for accessibility. The extent to which the content that would form the basis for an acceptable response to the prompt at hand and/or the awareness of which writing skills are required to shape an acceptable response should fall within the capabilities of the widest possible range of candidates.

Difficulty and accessibility are elements that factor into (a) the prompts, (b) the test takers, and (c) the scoring procedures used to rate the writing samples produced.

Prompts. Within the linguistic formulation of the prompt, there is clearly a wide range of room for misinterpretation, which must be closely guarded against to allow test takers to fully grasp what they are being asked to do. There seems no reason to make a prompt more difficult by choosing vocabulary or sentence structure likely to cause problems for large numbers of candidates. Conversely, one might argue that a test taker who interprets the direction to "give concrete examples" in a prompt that was specifically focused on cooperation versus competition by giving examples about cement has more than a linguistics

deficiency and also lacks discourse-level understanding. The recent call by the TWE Committee to experiment with glossing vocabulary items that might otherwise make a prompt opaque seems well-motivated and, if implemented, will provide good insight into how to minimize linguistic barriers to prompt decoding.

In addition to linguistic features, prompts also tap into the knowledge base of the test takers either through the way in which they are worded or through the presentation of material (orally or visually) that focuses and restricts the possible content the test taker can write about. In assessment instruments using "bare" prompts, that is, those that are presented without the framework of immediately available background material to apprehend and/or review, topics obviously range through a wide panoply of possible content areas: to argue for or against the government building an airport, factory, or hotel in one's community taps into completely different knowledge and content than to argue for or against requiring sports in the school curriculum, or to argue the superiority of newspaper, radio, or television news. While all of these topics can be presented in similar rhetorical frames, their content area concerns are more or less unique and are unlikely to overlap to any real degree. The test taker comes into the testing room and is presented with a topic that may immediately tap into a highly familiar issue, or be completely unfamiliar. With each test taker bringing his or her own entire range of personal knowledge of the world to the exam and then being required to draw upon a topic area that might fall at the periphery of that writer's knowledge, it seems particularly unfair to offer a test that yields one sample and provides no options. While it is certainly true that "it is absolutely obvious that for any given topic area, some candidates might have more knowledge or experience than others" (R. Kantor, ETS personal communication, 8/93), designing assessments that offer equally accessible prompts is extremely difficult because of the enormous range of variables that impact each individual writer's writing performance: Pollitt et al. (1985), for instance, list 32 characteristics *just within the textuality of the prompt itself* that impact on the difficulty of a prompt. Giving test takers choices on writing tests is a method that is often tried, but the research (described in detail in Hamp-Lyons, 1990) is ambiguous on the effects of such choices, and Hamp-Lyons and Mathias (1994) report research suggesting that supposedly equivalent prompts on the MELAB are not, in fact, equivalent. A similar note of caution must echo from the Golub-Reese et al. (1993) study of TWE prompts.

Lastly, prompt difficulty and accessibility are connected to the nature of the tasks the prompt asks the writer to perform. Critical questions to ask include: Are the tasks cognitively easy or difficult? Are the tasks stated implicitly or explicitly? If there is more than one task in a prompt, can all tasks be completed in the time allowed? Tasks here can be subdivided into the work required of the test taker at the cognitive level and the work required at the rhetorical level. The Hale et al. (1996) study classified all the prompts in its database along these two dimensions. While it is relatively difficult to define with precision the cognitive demands a particular writing task requires a writer to perform in order to create a satisfactory text, Hale et al. made the decision to bifurcate tasks into (a) those that were relatively undemanding and that called on the writers to retrieve, organize, or relate information versus (b) those that required more cognitive manipulation, ranging from the simpler demand of applying some known concepts to a new arena, to the more complex ability to synthesize and evaluate a range of materials, concepts, or ideas.[3] (An alternate perspective on cognitive demands can be found in Kirkland & Saunders, 1991.) At the rhetorical level, the Hale et al. study identified the mode of discourse required to address the writing assignment (i.e., narration, description, exposition, argument, or any possible combination of these) as well as the rhetorical

patterns the final product would be likely to demonstrate if the writer adhered to the demands of the assignment. Clearly, the development of a research agenda into prompt variables and prompt effects is a major task in itself, for there are so many unresolved questions that careful thought must be given to prioritizing these based on how much impact decisions about prompts will have on test takers' options and, therefore, on their performance. In addition, the interaction between decisions about task types and genres may affect choices and effectiveness of scoring. For instance, variations in genre have been shown to have a potential impact on judging and grading essays. Hake (1986) concludes that personal experience expository essays that incorporate narrative tend to be graded more objectively than pure narratives that describe personal experience. And even if the test prompts are not designed to elicit narrative, it is possible for some test takers to turn their responses into narrative.

The test takers. Another avenue of research to pursue relates to how the test takers view the difficulty and accessibility level of particular topics. Most writing tests have pre-testing procedures that include an evaluation of the viability of every potential topic for an operational administration. But while writing teachers and test administrators are routinely asked to comment on the prompts their students try out, the writers themselves are generally not asked to fill out questionnaire data or to participate in interview sessions to analyze their reactions to the topics they write about, which could be one way of gathering some critical information about how writers process and interpret prompts. Certainly, such research can produce inconclusive and problematic results. In a recent study of 20 students enrolled in a freshman composition course and writing to TWE-like prompts calling for either narrative or argument, Schaeffer (1993), using both questionnaires and interviews, asked students to assess their sense of whether or not they found a particular prompt difficult. Her findings indicate no correlation between students' perceived reaction to a prompt's difficulty and their ability to receive a passing score (4.0 or higher) on timed essays that were graded according to a 6.0 rubric similar to the TWE scoring guide. Similarly, Peretz and Shoham (1990), in a study with 177 EFL students, report that students' subjective evaluation of a text's difficulty did not correlate with their scores on multiple-choice comprehension tests using texts that were within and outside the students' disciplinary areas of study. Findings like these should cause us to hesitate to give test takers choices on writing tests that have only a single sample, because there is no evidence that student writers can make good choices for themselves. However, if we could learn what factors test takers use when judging whether a prompt will be difficult or easy to write on, we could design documentation for the test that would help writers think about TOEFL 2000 writing prompts in ways that might be more fruitful for the specific test environment.

Scoring procedures. Prompt difficulty is directly connected to scoring because raters will be asked to either strictly adhere to the ways in which test takers' writing samples instantiate particular features of the scoring guide (making a prompt more difficult) or to overlook and perhaps ignore instances of well-intentioned but misguided interpretations of the prompt (rendering it easier). A test of communicative competence should demand that test takers demonstrate their writing skills within the framework of an entire discourse package, upping the ante, so to speak, as regards the difficulty level vis-a-vis scoring. The discussions of scoring and of readers and reader training are relevant here.

Time

Several previous sections have discussed implications for test design in terms of the time spent in a writing test. Within the realistic constraints faced by TOEFL 2000, it is hardly possible to create a testing situation in which students can have space to fully exercise all processes; however, it seems important to provide sufficient time for students to make choices and decisions and to be able to follow them through. For example, on a two-hour writing test, it would be possible to have test takers do two or three different type tasks, none of which required a long, essay-style response, therefore giving them time to work through prewriting and revising/editing on at least one of the tasks, and more if they use time carefully. Some test takers would find the time more than they knew how to use; but two kinds of writers would have an advantage: intermediate to high intermediate writers who have been taught, or who intuitively possess, a good understanding of their own writing processes, would have the opportunity to use the time to actually apply their writing processes and strategies, and show themselves at their best. And writers who find timed testing situations particularly difficult would encounter less time pressure and (if time constraint is truly the locus of their test anxiety) would be able to spend time thinking through what they want to do before actually starting to write.

It would be inaccurate to think of a writing test that features two or three tasks in two hours as anything other than a speeded test, or to imagine that it can balance process with product, but it would certainly redress the usual imbalance to some degree. Studies showing that increasing test time for a writing task from 20 to 30 or from 30 to 45 minutes makes no difference to score and does not address the concerns of the rhetoric/composition community. For example, a study of different lengths of time for writing test tasks carried out by the Michigan State Department of Education (while one of the authors was a member of that body's Writing Advisory Committee) showed that at grades 4 and 7, only increasing time from 30 minutes to 2 hours made a difference; on the other hand, at those grade levels, a time increase from 30 minutes to 1 hour made no difference, and additional time made no extra difference. At grade 10, however, not only did increasing time by 2 hours make a difference, but further increasing it to two days (i.e., allowing an overnight rest period before returning to the writing test) also increased scores. It seems that for more advanced writers every minute of time can be put to good use.

Scoring Writing Samples

It is impossible to imagine a situation in which writing samples produced by test takers could be scored by computer, thus whatever shape the writing component of TOEFL 2000 takes, texts will be produced that must be read by people trained to apply scoring criteria in a principled, uniform, and systematic manner. Even if the score assigned to a particular writing sample is to be statistically factored into a single TOEFL score, raw scores assigned to the writing sample(s) must have a sufficient range of possibilities so that they represent and distinguish among the range of proficiency levels exhibited in the total sample pool. For TOEFL 2000 to have the highest possible validity as a measure of writing, it is important to design and control the way in which test takers' samples will be scored, and it is important to design and control prompts in which the range of variability from test form to test form is not significantly different. (This is just another example of the complicated way in which the many interlocking factors of test design might have economic impact for the program.)

Decisions about scoring procedures follow inevitably from decisions about prompts, text type, time allowed, and all the other prior contextual constraints that are applied. There are, however, some key aspects of scoring that must be considered early on in test design.

The first of these is, of course, reliability. We are becoming much more aware of the limitations of the traditional definition of reliability on writing tests as a coefficient based on two readers' scores (which, as Eliot, Plata, and Zelhart have illustrated [1990; pp. 88-89], can be above .90 even when readers never give identical scores), or even as a percentage of agreement between readers. Other factors work to lower the reliability of writing tests: variability of prompt difficulty which, as the prompt section has explained in detail, is a serious and so far unsolved problem with writing assessments; variations in writers' background knowledge; and so on. While all these fundamentally speak to issues of validity, we are now in possession of psychometric methods that enable us to separate out the effects of the different factors making up judgments about language performance. Before discussing these in greater detail, let us review exactly why this is a critical issue.

The six-point scoring system used to rate many writing assessment exams allows for 11 different scores (ranging from 1.0 to 6.0 at half-point intervals), which might be interpreted to mean that the test results account for 11 different gradations of proficiency. This view, however, is flawed because it does not take into account the manner in which the scores are actually assigned. For example, a given paper that can be seen as a high 4 or low 5 because it exhibits a true mixture of qualities present at both those points in the scale presumably deserves a true score of 4.5. However, depending on the actual readers in whose packets the paper appears, that individual paper can get a score of 4.0, if both readers tend to see it more "4-like" than "5-like," or a score of 5.0, if both readers tend to see it more "5-like" than "4-like." Only if the paper is by happenstance read by a more harsh reader paired with a more lenient reader — the former scoring it a 4 and the latter giving it a 5 — will it receive its "true" score of 4.5. Since the current reading procedures used by many programs that use a six-point holistic or general impression scoring scale do not control for this, the gradations between final assigned scores of 4.0, 4.5, and 5.0 (to take just one spread possible) cannot be seen to be inherently meaningful since the conditions under which a true score can be reliably ascertained do not exist. There are, however, ways to overcome this problem.

In the last decade, latent trait theory, also referred to as item response theory, has been applied to language testing data and been found very useful. Item response theoretic methods are becoming widely used in language testing and in educational measurement. Partial credit models of item response theory are now popularly used to analyze data where there are more than simple "right/wrong" (i.e. dichotomous) answers (the rating scales we use for judging writing are classic instances of partial credit scoring). One-parameter Rasch analysis, particularly the multiple-facet Rasch analysis used by the computer program FACETS, has made it possible for us to identify raters who are more or less severe, prompts that are more or less difficult, as well as writers who are more or less proficient. FACETS also makes it possible for us to create other analytic categories that can help us understand potential bias in a writing test, for example, by ethnicity or language background. FACETS could be a useful tool in many of the necessary research studies to be undertaken in the development of the writing test for TOEFL 2000, but in addition we should also think about whether FACETS (or other similar programs that will no doubt come along in time) can be used in the analysis of score data as it is produced so that test takers' scores

can be reported on a scale that has been responsive to the difficulty of the prompt and the severity of the rater as well as the ability of the writer.

The limitation placed on FACETS and all item response theoretic approaches in the typical writing assessment context is the lack of crossover or overlap in the data. On a typical such exam, each writer writes only one text — there is no crossover of prompts to enable us to see whether the student's writing ability varies according to prompt. At a typical reading, each rater scores hundreds of essays, but all on the same prompt — there is no crossover of raters to enable us to see whether the rater's scoring pattern varies according to prompt. We shall not get very far with research into prompt specifications, prompt equating, rater training, and impact of rating scale(s) for TOEFL 2000 unless we develop research and operational plans that build in crossover. With FACETS this does not have to be fully orthogonal — in fact, it can be quite limited — but it must exist.

The development of multiple-trait item response models has led some people to ask whether we can then do away with second readings of essays. If we can know the relative harshness or leniency of a rater, can we not adjust her or his scores against the model without checking them against those of a second rater? This might be psychometrically viable in scoring mainstream writing (though it will always be severely criticised by the first-language rhetoric/composition community). But in the L2 context writers are so varied in their backgrounds that it would be unreasonable to assume that a computer program could find a large enough sample of essays written by comparable writers and scored by the same reader against which to estimate the harshness or leniency of *this* rater in *this* specific nexus of circumstances. Hake (1973), using an early predecessor of FACETS, found that raters from different backgrounds exhibited different patterns of harshness or leniency for different prompts *and for writers from different backgrounds*. Each rater would need to rate so many trial essays in order to establish a personal baseline that it would be simpler to do double rating as usual and use FACETS with a more manageable set of variables.

The second aspect of scoring to be considered is validity. We are always caught in a bind over reliability and validity. The psychometric wisdom is that no test can be more valid than it is reliable — if scores are not stable or consistent, then they are essentially meaningless and cannot be valid. But the pendulum swings both ways; if a test is not valid, there is no point in its being reliable, since it is not testing any behavior of interest. This is, of course, a modern view of validity based on the work of, for example, Anastasi (1982), Cronbach (1988), and Messick (1988), which emphasizes construct validity. It is concerns about construct validity that drive our discussion of prompts, of the assessment of process as well as of product, and here, of scoring procedures.

Beyond Holistic Scoring

Although ETS has spearheaded the refinement of the technique into a very efficient and accessible tool, traditional holistic scoring is still characterized by impression marking and the use of multiple raters to compensate for interrater unreliability (Cooper, 1984). While there are a number of serious problems with this form of holistic scoring in any context (White, 1993), these problems are especially serious in ESL writing assessment contexts. In speeded, impressionistic holistic scoring, many raters make judgments by responding to the surface of the text and may not reward the strength of ideas and experiences the writer

discusses. It is difficult for readers making single judgments to reach a reasonable balance among all the essential elements of good writing. In ESL contexts, a detailed scoring procedure requiring the readers to attend to the multidimensionality of ESL writing may ensure more valid judgments of the mix of strengths and weaknesses often found in ESL writings. In the scoring of the IELTS writing, for example, a multiple-trait scoring instrument is used, enabling raters to judge essays on features found to be salient to raters during trial scoring sessions. Higher order features of writing such as the presence and quality of evidence are used to make judgments of one task type that elicits writing that uses evidence, while in a different task type this feature was not salient and was therefore not built into the scoring procedure. Instead, different features found to be particularly salient in that different task context, such as cohesion and coherence, were judged. A similar process was used in the development of the original scoring procedure for the ELTS writing (Hamp-Lyons, 1987). Each of these features generates its own score and each score is separately reported. Not only does this clarify for test takers, raters, and score consumers what writing skills are necessary to complete each writing task effectively, it also helps raters balance their judgments of characteristic ESL features of writing, principally a high frequency of low-order sentence grammar problems, against higher order elements of the writing, such as the use of evidence, rhetorical control, and so on. These ideas are discussed in detail in Hamp-Lyons, 1991b and in press-b, and have been implemented in several tests — the ELTS, IELTS, **access:**, and the University of Michigan writing assessment for undergraduates — and have been adapted by a number of institutions in the United States and elsewhere.

In addition to the theoretical difficulties with holistic scoring, the nature of holistic judgments presents a practical weakness, since scores generated in this way cannot be explained either to the other readers who belong to the same assessment community and who are expected to score reliably together, or to the people affected by the decisions made through the holistic scoring process — the test takers/writers, their parents and teachers, and their academic counselors. Not only is diagnostic feedback out of the question, there is even the danger that a call for accountability in testing will be met only with the response, "It's a '4' because I say so" — a response each of us is uncomfortable with both theoretically and practically. While speeded holistic readings typically generate fairly high classical reliability (around .85) and multiple-trait readings generate classical reliabilities that are only slightly higher, it is not understood how one scoring procedure or another affects the kind of reading the rater gives the text, that is, how the scoring procedure affects the validity of the judgments the raters make. O'Loughlin (1993) found suggestions of differences in the kinds of judgments raters were making in his study of traditional holistic scoring versus analytic scoring (his analytic instrument was a slight modification of Hamp-Lyons' multiple-trait instrument for one of the IELTS writing tasks). Research to discover if and how different scoring procedures affect scores for ESL essays, and if scores vary according to the kinds of raters doing the scoring, is necessary.

Readers and Reader Training

Whatever type of scale or rating procedure is determined appropriate for the writing samples that will be produced by TOEFL 2000 examinees, we also need to consider who will be employed to do the actual reading of the samples and how those readers will be trained. Where non-ESL writing specialists are used to score ESL writing samples, attention must be paid to providing these specialists with opportunities to distinguish global errors (errors that interfere with communication) from local errors (errors that may be

irritating but do not interfere with comprehension). Research studies regarding the writing of native versus nonnative English speakers and ratings by teachers of native versus nonnative English speakers have provided conflicting results. In a pilot project that explored the ratings assigned to essays produced for the English Placement Test administered by the province of British Columbia in 1979, McDaniel (1985) concluded that raters did not respond "the same way towards essays written by ESL and non-ESL students," (p. 15). More recently, Brown (1991) found that while essays written by University of Hawaii students enrolled in either English composition or ESL composition classes and rated by ESL and non-ESL teachers received similar scores, the teachers did not agree on what components of the student texts could be classified as either the worst or the best features. Sweedler-Brown (1993) also concluded that raters not trained to work with ESL students tend to judge essays on different criteria than they use to judge essays written by native English speaking students. Vann, Meyer, and Lorenz (1984) found that faculty response to ESL students' errors varied in many ways; O'Loughlin (1993) found that under some conditions English and ESL-trained raters rated ESL essays differently.

Most ESL writing tests use only ESL-trained readers in scoring, but the present TWE test is an exception. Probably due to its great size, the TWE test uses a group of raters who are predominantly English-trained and not ESL-trained, and current rater training pays no specific attention to how ESL writing features affect the TWE essays. While the weight of the studies reported above suggests that ESL-trained readers are likely to give more valid (and reliable) readings of ESL essays, ETS has not undertaken a study to see if non-ESL and ESL teachers react to the same features of an essay in arriving at a holistic TWE score since the Carlson, Bridgeman, Camp, and Waanders (1985) study. This study found only moderate agreement between the scores of the two kinds of readers, and this is an important issue that deserves to be revisited.

In addition to identifying more precisely whether raters who have not received ESL training can properly be used as raters of ESL writing, and if so, how best to jointly train readers who are trained in ESL with those who are not, other variations in rating procedures should be considered. For instance, there is little evidence to indicate whether readers can read more than one prompt during a single scoring period without negative effects on reliability. While there are many writing tests in which readers read only a single prompt, there are also many others in which raters read a mix of prompts. Usually, the impact of the different approaches goes unresearched because either one or the other approach is used and no research studies are set up to compare the two. By cycling raters to the judgment of different prompts over a two- to three-day rating period, data would be available about rater-prompt interaction effects, using Multivariate Analysis of Variance (MANOVA) or (preferably) Item Response Theory (IRT).

Score Reporting and Score Use

While TOEFL/TWE has never been intended for use in diagnostic contexts, one way to encourage the wider acceptance of the future TOEFL 2000 writing test component is to give it a diagnostic function and structure. Schools with only a small service unit for foreign students could be relieved of the need to design and administer their own writing test and instead require students to produce their diagnostic report from the TOEFL 2000, or they could have students request a report to be sent from ETS for an additional fee. For placement decisions on individual campuses, ESL writing advisers can establish cutoffs so that test scores correspond to the perceived gradations in individual courses. For example, a program with four levels of ESL course scores could be reduced to a four-step scale that would equate to something like 1 = "proficiency below the minimal level required for entering students;" 2 = "performs at the level of course X;" 3 = "performs at the level of course Y;" and 4 = "skills merit exemption from further coursework," with the TOEFL 2000 writing score equivalents for each of the four steps clearly stated. The score report can be created in such a way that it could be translated to a wider or narrower range as need dictates. While a more detailed scoring procedure is more expensive in terms of readers' time and database management, the additional utility of the scores produced would compensate for this (see below, "Cost").

The development of scoring guides and rating scales for all writing tests is a skilled, complex, and time-consuming activity. In the ESL context it is both more complex and more critical, since the way that ESL writing performance expectations are stated in the rating scale will both shape readers' judgments and impact the teaching of writing around the world. Jacobs, Zingraf, Wormuth, Hartfiel, and Hughey (1981) provide a strong example of a carefully developed and implemented ESL writing assessment with special attention paid to the development of a scoring procedure. We can benefit also from studying the solutions found by other large-scale, international writing tests and tests that incorporate writing with other skills, in particular the ELTS and IELTS. Both these tests used a type of modified analytic scoring, and the ELTS had different traits for scoring two different kinds of prompts (Hamp-Lyons, 1987). A close study of these and other tests (MELAB, **access**:, TEEP, and so on) would be a fairly simple but very useful review. We can also look to scoring procedures that are not merely holistic but ask raters to attend to specific features in the writing that seem critical for ESL writers to control, such as cohesive ties (see Jafarpur, 1991), or those that are more directly connected to the type of task the writer is asked to address (see Connor, 1991) in a modified primary trait assessment.

To address these and other problems in test scoring procedures, a number of decisions must be made involving potential trade offs between factors that might improve reliability and validity but that might decrease cost effectiveness and considerably lengthen the amount of time required to score each writing sample produced for the test. The components that must be factored into development of scoring guidelines for the writing samples to be produced on TOEFL 2000 include the criteria by which essays will be rated, the ways in which readers will be trained to apply those criteria, and the procedures that will be utilized to assure that the same guidelines are applied to each administration of the test. Criteria for scoring the essays should be determined by both the purposes to which the test scores will be used and the appeal to the qualities of proficient or good writing that appear to be uniformly relevant to English prose of the genre(s) the writing samples require. Research should also address whether measures of reading speed and inter-rater agreement suffice to measure the construct validity of raters' judgments.

Closing Thoughts

Costs, Practicality, and Washback

Attention is currently focused on issues of washback in language testing, and indeed arguments about washback had a positive influence on the decision made by ETS to implement TWE as part of the TOEFL program. A test of writing, it was argued, would make it more likely that writing would be taught in intensive English and other English preparation programs, in North America and overseas. In planning the writing component of TOEFL 2000, whether as a single-skill test or within an integrated model, it is vital to hold onto those arguments. If TOEFL is to spend money on a new test and include writing in that test, it is sensible to spend these resources wisely in order to create a test that will really tell us something about the writing strengths (and perhaps weaknesses) of the test takers, and that will also have potential multiple uses. Like the photo-imaging hardware and software that was created to include the test taker's photograph with the score report, technology might be reconfigured so that the TOEFL 2000 writing test can ultimately best serve score users through its potential to transmit uncompromisingly genuine samples of a test taker's writing. As a testing procedure becomes more complex, it also becomes more costly. However, some of the cost of assessment can be won back in increased information. If scores are useful for more than one purpose, they may be requested (and paid for) more often. If scores are more useful, the cost for a test can be higher. Studies should determine whether the additional information that would be available through implementing some of the suggestions about scoring methods and score reporting would be marketable to specific categories of actual and potential score users.

In addition there is the whole area of test impact, which is outside the range of this report. Nevertheless, it is clear that not only must TOEFL 2000 contain writing test tasks (alone or integrated with tasks testing other skills), but it must do so in ways that will be seen by applied linguists, admissions officers, college faculty, teachers teaching TOEFL 2000 preparation courses, and most of all, the test takers to be connected to the expectations of the college communities the test takers will enter. The use of single-sample context-stripped prompts with 30 minutes to write responses no longer has validity in the eyes of the ESL composition community. To date, however, there is little hard evidence to back up claims about the beneficial washback effects of direct-writing tests compared to multiple-choice tests. Indeed, Alderson and Hamp-Lyons (1993) suggest that while teaching TOEFL preparation may lead teachers to change the content of what they teach, there is no evidence so far to suggest that their teaching methods are changed in the TOEFL preparation context. Clearly, washback is a more complex concept than simple claims suggest, and washback studies could usefully inform decisions about which of many ideal innovations will be fruitful for TOEFL 2000.

Work of Test Development

In the actual practice of constructing writing prompts for testing purposes, the TOEFL 2000 test developers must work with several integrally related and complex variables: the modalities in which the test will be framed (print, oral, visual); the wording of both the prompt and the instructions surrounding the prompt, both of which involve a lot of linguistic variables; the subject matter for the test question or questions (content variables); the rhetorical specifications, if any, embedded in the task; and the level of cognitive demand to be placed on the test taker. Each of these variables interacts with the others to shape the task, such that manipulation of one variable might require manipulation of another, which then no

longer holds true in quite the same way once the first variable is given a particular instantiation. Thus, the work of the test development committee will be particularly complex. The development of the scoring procedure must go hand-in-hand with the development of the prompt specifications, template(s), and exemplars. Once this work is completed, and the development of a large pool (or, we hope, bank) of prompts begins, the test development committee should keep the scoring criteria at the forefront of their minds to guide their selection of prompts that will generate writing that can be validly scored by faithful adherence to the scoring guide.

Whatever shape the TOEFL 2000 writing component takes, it will require a highly regulated set of procedures to develop the prompts that will be used to elicit writing. If, for example, the decision is made to use reading material as the stimulus to create the writing task(s) to which students will respond in writing, the principles under which authentic raw reading materials (texts) are to be selected must be precisely identified and strictly controlled. Similarly, if the decision is made to use oral material, such as a videotaped or audiotaped lecture or a discussion between two or more speakers, the precise parameters for selecting or creating this material must be spelled out, starting with whether oral material needs to be scripted or will derive from authentic academic language. If a combination of written and verbal starting material is going to form the basis for the writing prompt(s), that further expands the number of factors that must be identified, controlled, and equated from test form to test form in order to assure the validity of the exam. In addition to the source material that could be used to shape the writing component of the exam, there will remain a specific question or questions (or a choice of questions) to which the test takers must respond in continuous prose to produce a writing sample, which will be scored for levels of proficiency. For regardless of the modality of the originating stimulus, test takers will still be asked to write to a prompt (or prompts) designated in the test itself and not self-designed by the test takers. The more variables there are in the materials that are used to lead into the prompt, the more oversight must be factored into the work involved in the development of the prompts. And if the TOEFL 2000 writing test is to be integrated with other skills, periodic meetings of all test development committees together will be needed.

In large-scale language tests, such as TOEFL 2000, Kroll and Reid (1994) point out that the constraints for designers of writing tasks are particularly daunting because there is no margin for error: the prompts are written for an enormous unknown audience, and there will be no opportunity to make on-site adjustments or revisions. Multi-item tests, such as IELTS and **access:**, have prompts at a hierarchy of difficulty levels, which points to another reason for the writing component of TOEFL 2000 to be a multi-item test.

Recommendations

From the mass of specifics we have covered in this overview, we find the following to be especially worth considering:

1. The arguments for a test with more than one task seem overwhelming.

2. The same is true for tasks of more than one task type.

3. In a multiple-sample test, the weight of evidence suggests that the test takers should be able to choose at least one element of the test.

4. Evidence overwhelmingly suggests that graduates and undergraduates should be offered different tests.

5. Before raters not trained in ESL are used, the effects of using ESL- versus English-trained raters should be carefully researched; raters should be trained with the specifics of ESL writing in mind.

6. Research should focus on the costs and benefits of a more fully articulated scoring procedure.

7. Information about an individual rater's harshness or leniency should be used as a factor in establishing each writer's score.

8. Multiple forms of score reporting should be considered.

Footnotes

[1]The ELTS provided six disciplinary modules (described in detail in Hamp-Lyons, 1991b): Social Studies, Physical Sciences, Life Science, Technology, Medicine, and General Studies.

[2]"Pool" is a store of prompts to be used subsequent to their storage in the pool; "bank" is a store of prompts, known to be equivalent, that can be reused.

[3]While recognizing that cognitive demands range across a broad spectrum of difficulty level, the researchers made the decision to use only a two-way distinction because of the difficulty of determining the exact nature of each assignment being analyzed.

References

Ackerman, J. M. (1990). Students' self-analyses and judges' perceptions: Where do they agree? In L. Flower, V. Stein, et al. *Reading to write: exploring a cognitive and social process* (pp. 96-115). New York: Oxford University Press.

Ackerman, J. M. (1993). The promise of writing to learn. *Written Communication, 10,* 334-30.

A language for life. (1975). *The Bullock Report.* London: Her Majesty's Stationer's Office.

Alderson, J. C. (1981). Report of the discussion of general language proficiency. In J. C. Alderson & A. Hughes (Eds.), *Issues in language testing: ELT Documents 111* (pp. 187-193). London: The British Council.

Alderson, J. C., & Hamp-Lyons, L. (1993, August). *TOEFL preparation courses: A study of washback.* Paper presented at the Association Internationale Linguistique Applique, Amsterdam: The Netherlands.

Anastasi, A. (1982). *Psychological testing* (5th edition). London: Collier Macmillan.

Auerbach, E. R. (1986). Competency-based ESL: One step forward or two steps back? *TESOL Quarterly, 20*(3), 411-429.

Bartholomae, D. (1985). Inventing the university. In M. Rose (Ed.), *When a writer can't write: Writer's block and other composing process problems* (pp. 134-165). New York: Guilford.

Becher, T. (1987). Disciplinary discourse. *Studies in Higher Education, 12*(3), 261-274.

Belanoff, P., & Dickson, M. (Eds.). (1991). *Portfolios: Process and product.* Portsmouth, NH: Boynton/Cook Heinemann.

Berlin, J. A. (1982). Contemporary composition: The major pedagogical theories. *College English, 44,* 765-777.

Berlin, J. A. (1987). *Rhetoric and reality: Writing instruction in American colleges, 1900-1985.* Carbondale, IL: Southern Illinois University Press.

Berthoff, A. (1978). Tolstoy, Vygotsky, and the making of meaning. *College Composition and Communication, 29,* 249-255.

Bridgeman, B., & Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate students* (TOEFL Research Report #15). Princeton, NJ: Educational Testing Service.

Britton, J., Burgess, T., Martin, N., McLeod, A., & Rosen, H. (1975). *The development of writing abilities (11-18).* London: Macmillan Education.

Brossell, G. (1983). Rhetorical specification in essay examination topics. *College English, 45*, 165-174.

Brookes, G., Markstein, L., Price, S., & Withrow, J. (1992, March). *Portfolio assessment at Manhattan Community College, City University of New York.* Paper presented at the 26th TESOL Convention, Vancouver, Canada.

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

Bruffee, K. A. (1986). Social construction: Language and the authority of knowledge; a bibliographical essay. *College English, 48*, 773-790.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M. A. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 45-78). Cresskill, NJ: Hampton Press.

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.

Carlson, S., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English* (TOEFL Research Report #19). Princeton, NJ: Educational Testing Service.

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

Chan, S. (1990). *Language in a small business management class at the Chinese University of Hong Kong.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Chomsky, N. (1957). *Syntactic structures.* Cambridge, MA: Michigan Institute of Technology Press.

Chomsky, N. (1966). *Aspects of a theory of syntax.* Cambridge, MA: Michigan Institute of Technology Press.

Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing, 1*, 70-81.

Cohen, A. D., & Cavalcanti, M. C. (1990). Feedback on compositions: Teacher and student verbal reports. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 155-177). New York: Cambridge University Press.

Collignon, F. F. (1993). Reading for composing: Connecting processes to advancing ESL literacies. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp. 258-273). Boston, MA: Heinle and Heinle.

Connor, U. (1991). Linguistic/rhetorical measures for evaluating ESL writing. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 215-225). Norwood, NJ: Ablex.

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

Connors, R. (1981). The rise and fall of the modes of discourse. *College Composition and Communication, 32*, 444-455.

Cooper, C. R. (1984). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-32). Urbana, IL: National Council of Teachers of English.

Cronbach, L. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning, 39*, 81-141.

Eliot, N., Plata, M., & Zelhart, P. (1990). *A program development handbook for the holistic assessment of writing*. Lanham, NY: University Press of America.

Faigley, L. (1986). Competing theories of process. *College English, 48*, 527-542.

Flower, L. (1985). *Problem-solving strategies for writing* (2nd. edition). San Diego: Harcourt Brace Jovanovich.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*, 365-387.

Flower, L., & Hayes, J. R. (1984). Images, plans and prose: The representation of meaning in writing. *Written Communication, 1*, 120-160.

Flower, L., Stein, V., Ackerman, J., Kantz, M. J., McCormick, K., & Peck, W. C. (1990). *Reading to write: Exploring a cognitive and social process.* New York: Oxford University Press.

Gee, J. P. (1989). What is literacy? *Journal of Education, 171*(1), 18-25.

Golub-Smith, M., Reese, C., & Steinhaus, K. (1993). *Topic and topic type comparability on the Test of Written English* (TOEFL Research Report #42). Princeton, NJ: Educational Testing Service.

Greenberg, K. L. (1986). The development and validation of the TOEFL writing test: A discussion of TOEFL Research Reports 15 and 19. *TESOL Quarterly, 20*(3), 531-544.

Hake, R. (1973). *Composition theory in identifying and evaluating essay writing.* Unpublished doctoral dissertation, University of Chicago.

Hake, R. (1986). How do we judge what they write? In K. L. Greenberg, H. S. Weiner, & R. S. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 153-167). New York: Longman.

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 #44). Princeton, NJ: Educational Testing Service.

Hamp-Lyons, L. (1986). *Assessing writing in academic settings.* Unpublished doctoral dissertation, University of Edinburgh, Scotland.

Hamp-Lyons, L. (1987). Performance profiles for academic writing. In K. Bailey, R. Clifford, & E. Dale (Eds.), *Selected papers from the Language Testing Research Colloquium.* Monterey, CA: Defense Language Institute.

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

Hamp-Lyons, L., & Reed, R. (1990). Developing the Michigan Writing Assessment: Report to the College of Liberal Sciences and Arts. Mimeo: English Composition Board. Ann Arbor, MI: University of Michigan.

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

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

Hamp-Lyons, L. (1993, April). Components of portfolio evaluation: ESL data. Paper presented at the meeting of the American Association of Applied Linguistics, Atlanta, Georgia.

Hamp-Lyons, L. (in press-a). Exploring bias in essay tests through student interviews. In J. Butler, J. Guerra, & C. Severino (Eds.), *Writing in multicultural settings.* New York: Modern Language Association.

Hamp-Lyons, L. (in press-b). The challenges of second language writing assessment. In E. White, W. Lutz, & S. Kamusikiri (Eds.), *The practice and politics of writing assessment.* New York: Modern Language Association.

Hamp-Lyons, L., & Mathias, S. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3,* 1.

Hayes, J. R., & Flower, L. (1983). Uncovering cognitive processes in writing: An introduction to protocol analysis. In P. Mosenthal, L. Tamar, & S. A. Walmsley, (Eds.), *Research in Writing* (pp. 206-220). New York: Longman.

Herrington, A. (1986). Studying writing in academic contexts: The view from within our classrooms. *PALM (Papers in Applied Linguistics, Michigan), 2*(1).

Herrington, A., & Moran, M. (1992). *Writing, teaching and learning in the disciplines.* New York: Modern Language Association.

Horowitz, D. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20,* 445-462.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth, England: Penguin.

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

Jafarpur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System, 19,* 459-65.

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

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. M. (1991a). Faculty assessment of ESL student literacy skills: Implications for writing assessment. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 167-179). Norwood, NJ: Ablex.

Johns, A. M. (1991b). Interpreting an English competency examination: The frustrations of an ESL science student. *Written Communication, 8,* 379-401.

Johnson, K. (1981). *Communicate in writing.* London: Longman.

Johnson, K. (1983). Communicative writing practice and Aristotelian rhetoric. In A. Freedman, I. Pringle, & J. Yalden (Eds.), *Learning to write: First language, second language* (pp. 247-257). London: Longman.

Kantz, M. (1990). Helping students use textual sources persuasively. *College English, 52,* 74-91.

Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly, 25*, 105-121.

Knoblauch, C. H. (1988). Rhetorical considerations: dialogue and commitment. *College English, 50*, 125-140.

Kroll, B. (1990). Introduction. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 1-5). New York: Cambridge University Press.

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

Kuhn, T. S. (1970). *The structure of scientific revolutions (2nd edition)*. Chicago: University of Chicago Press.

Labov, W. (1969). *The logic of non-standard English.* Urbana, IL: National Council of Teachers of English.

Langer, J., & Applebee, A. (1986). *How writing shapes thinking: A study of teaching and learning.* Urbana, IL: National Council of Teachers of English.

Leki, I. (1991). Twenty-five years of contrastive rhetoric: Text analysis and writing pedagogies. *TESOL Quarterly, 25*, 123-143.

Leki, I. (1992). *Understanding ESL writers: A guide for teachers.* Portsmouth NH: Boynton/Cook Heinemann.

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.

McDaniel, B. A. (1985, March). Ratings vs. equity in the evaluation of writing. Paper presented at the 36th Annual Conference on College Composition and Communication. Minneapolis, MN.

McGinley, W. (1992). The role of reading and writing while composing from sources. *Reading Research Quarterly, 27*, 227-50.

McKenna, E. (1987). Preparing foreign students to enter discourse communities in the U.S. *English for Specific Purposes, 6*(3): 187-202.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Murphy, S., & Ruth, L. (1993). The field testing of writing prompts reconsidered. In M. A. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 266-302). Cresskill, NJ: Hampton Press.

O'Donnell, W. R. (1968). *An investigation into the role of language in a physics examination.* (Moray House Publications Monograph, No. 7). Edinburgh, Scotland: Oliver and Boyd.

O'Loughlin, K. (1993, August). The assessment of writing by English and ESL teachers. Paper presented at the International Language Testing Research Colloquium, Cambridge, England.

Pennycook, A. (1989). The concept of method, interested knowledge, and the politics of language teaching. *TESOL Quarterly, 23*(4), 589-618.

Peretz, A. S., & Shoham, M. (1990). Testing reading comprehension in LSP: Does topic familiarity affect assessed difficulty and actual performance? *Reading in a Foreign Language, 7*, 447-55.

Perl, S. (1979). The composing processes of unskilled college writers. *Research in the Teaching of English, 13*, 317-336.

Pollitt, A., Hutchinson, C., Entwhistle, N., & De Luca, C. (1985). *What makes exam questions difficult?* Edinburgh: Scottish Academic Press.

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

Raimes, A. (1983). Anguish as a second language: Remedies for composition teachers. In A. Freedman, I. Pringle, & J. Yalden (Eds.), *Learning to write: First language, second language* (pp. 258-272). London: Longman.

Raimes, A. (1985). What unskilled writers do as they write: A classroom study. *TESOL Quarterly, 19*, 229-258.

Raimes, A. (1990). The TOEFL Test of Written English: Causes for concern. *TESOL Quarterly, 24*(3), 427-442.

Raimes, A. (1991). Out of the woods: Emerging traditions in the teaching of writing. *TESOL Quarterly, 25*(3), 407-430.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric point of view. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191-210). New York: Cambridge University Press.

Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing, 2*, 79-107.

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

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

Savignon, S. J. (1991). Communicative language teaching: The state of the art. *TESOL Quarterly, 25,* 261-277.

Schaeffer, C. (1993). *Pragmatic sequencing: A comparative case study of a freshman composition class in English.* Unpublished master's thesis, California State University, Northridge.

Shen, F. (1989). The classroom and the wider culture: Identify as a key to learning English composition. *College Composition and Communication, 40,* 459-466.

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.

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College Composition and Communication, 31,* 378-388.

Spivey, N. N. (1990). Transforming texts: Constructive processes in reading and writing. *Written Communication, 7,* 256-287.

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

Swales, J. M., & Feak, C. (1994). *Academic Writing for Graduate Students.* Ann Arbor: University of Michigan Press.

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

Valdez, G. (1991). *Bilingual minorities and language issues in writing: Toward profession-wide responses to a new challenge* (Technical Report No. 54). Berkeley, CA: Center for the Study of Writing.

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

Walvoord, B. (1986). *Helping students write well: A guide for teachers in all disciplines.* Urbana, IL: National Council of Teachers of English.

Weir, C. (1984). *Identifying the language problems of overseas students in tertiary education in the United Kingdom.* Unpublished doctoral dissertation. University of London.

White, E. (1993). Holistic scoring: Past triumphs, future challenges. In M. M. Williamson & B. A. Huot, (Eds.), *Validating holistic scoring for writing assessment* (pp. 79-108). Cresskill, NJ: Hampton Press.

Wolcott, W. (1987). Writing instruction and assessment: The need for interplay between process and product. *College Composition and Communication, 38*(1), 32-46.

Yancey, K. (1992). *Portfolios in the writing classroom.* Urbana, IL: National Council of Teachers of English.

Young, R., Becker, A., & Pike, K. (1970). *Rhetoric: Discovery and Change.* New York: Harcourt, Brace, Jovanovich.

Zamel, V. (1983). The composing processes of advanced ESL students: Six case studies. *TESOL Quarterly, 17*(2), 165-188.

Cover Printed on Recycled Paper

58701-14589 • Y37M.75 • 253700 • Printed in U.S.A.