TOEFL®

Monograph Series

MS - 16
APRIL 2000

TOEFL 2000 Framework: A Working Paper

Joan Jamieson
Stan Jones
Irwin Kirsch
Peter Mosenthal
Carol Taylor

ETS™ Educational Testing Service

TOEFL 2000 Framework:
A Working Paper

Joan Jamieson
Stan Jones
Irwin Kirsch
Peter Mosenthal
Carol Taylor

Educational Testing Service
Princeton, New Jersey
RM-00-3

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

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

E-mail: toefl@ets.org

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

Foreword

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

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

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

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

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

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

TOEFL Program Office
Educational Testing Service

Abstract

This paper lays out a preliminary working framework for the development of the TOEFL 2000 test. The goal of this first working framework is to guide the development of more specific frameworks and research agendas for the assessment of reading, writing, listening, and speaking, both as independent and integrated modalities.

The monograph is organized into five major parts. The first part presents a general introduction to the goals and key components of the project. The second part presents the historical background and work of the project leading to the development of this framework. The third and fourth sections present our conceptualization of a working framework that includes identifying the test domain, organizing the test domain, identifying task characteristics, identifying and operationalizing the variables, validating the variables, and building an interpretive scheme. The paper concludes with a discussion of the plans for proceeding with the work of the project.

Keywords: communicative language ability, framework, task characteristics, grammatical features, pragmatic features, discourse features

Acknowledgments

This working document could not have been completed without the support and contributions of many individuals across ETS and professional colleagues from the fields of applied linguistics and language teaching and testing. The current work has been shaped by the literature on communicative competence, the foundational efforts of the TOEFL 2000 project, and what has been learned about building empirically derived test frameworks from the adult literacy assessments developed at ETS.

Some, because of their special contribution, deserve mention. In particular, Frances Butler, Carol Chapelle, Dan Douglas, Daniel Eignor, April Ginther, William Grabe, and Robert Kantor provided extensive, critical reviews during the formative stages of the framework's development. We also benefitted from discussions with the TOEFL Committee of Examiners and Research Committee[1] on an earlier version of the paper. Finally, we wish to express gratitude to Lynn Jenkins whose careful editorial review improved the clarity of our text and presented our thoughts as one voice.

[1] When this framework paper was developed, the TOEFL Committee of Examiners and Research Committee were distinct committees. In fall 1997 these committees merged as a single committee, the TOEFL Committee of Examiners.

Table of Contents

Page

1. Introduction 1

2. Background of TOEFL 2000 Project 3

3. Conceptualizing a Framework 7

4. The Framework: Work to Date 9
   - 4.1 Identifying the Test Domain 10
   - 4.2 Organizing the Test Domain 12
   - 4.3 Identifying Task Characteristics 13
   - 4.4 Identifying and Operationalizing the Variables 14
     - 4.4.1 Situation 14
     - 4.4.2 Text Material 16
     - 4.4.3 Test Rubric 20
   - 4.5 Validating the Variables 24
   - 4.6 Building an Interpretive Scheme 31

5. Plans for Proceeding 39

References 41

Appendices
- Appendix A TOEFL Monograph Series 51
- Appendix B Description of the "Purpose" Variable 53
- Appendix C Grammatical Features 54
- Appendix D Rhetorical Properties 56
- Appendix E Text Structure Properties 58
- Appendix F Description of the "Type of Information" Variable 63
- Appendix G Description of the "Type of Match" Variable 66
- Appendix H Description of the "Plausibility of Distractors" Variable 72

List of Tables

		Page
Table 1	Analysis of TOEFL CBT Reading Tasks Using Variable Descriptors and Values	29
Table 2	Analysis of TOEFL CBT Listening Tasks Using Variable Descriptors and Values	30
Table 3	Internal Consistency of Variable Scores by Level for Prose Tasks in Five Adult Literacy Surveys	35
Table 4	Constructs Underlying Prose Task Difficulty and Examinee Proficiency by Level in Five Adult Literacy Surveys	36
Table C1	Text Features	54
Table C2	Vocabulary Features	55
Table D1	Rhetorical Types	56
Table D2	Interaction Types (Adjacency Pairs)	57
Table E1	Documents As Lists	58
Table E2	Graphs, Schematics, Maps, Forms	60
Table E3	Prose Text Structures	61
Table E4	Components of Interaction	62

List of Figures

		Page
Figure 1	Components of the TOEFL 2000 project	1
Figure 2	The structure of the TOEFL 2000 framework	9
Figure 3	A model of task characteristics	13
Figure 4	A model of task characteristics: situation	14
Figure 5	A model of task characteristics: text material	16
Figure 6	A model of task characteristics: test rubric	20
Figure 7	A model of prose and document processing in reading	26
Figure 8	"What Enabled Early Humans to Control the Use of Fire": sample text	27
Figure 9	The relation between task difficulty and examinee ability on the same IRT scale	32
Figure 10	Levels for the Chinese test mapped onto the IRT scale display	33
Figure G1	A text used to illustrate the "type of match" and "plausibility of distractors" variables	67
Figure G2	The additive scoring rubric used to characterize the difficulty of processing strategies and related conditions of use	70

1. Introduction

This paper presents a "working framework" for the development of a new computer-based Test of English as a Foreign Language, referred to hereafter as the TOEFL 2000 test. This new test will take into account recent literature on communicative competence, as well as previous work by TOEFL consultants, committee members, and staff.

To ensure the project's success, those creating the TOEFL 2000 test must consider three interdependent components that we believe are essential to operational testing programs: the constituencies, constraints, and framework (see Figure 1).

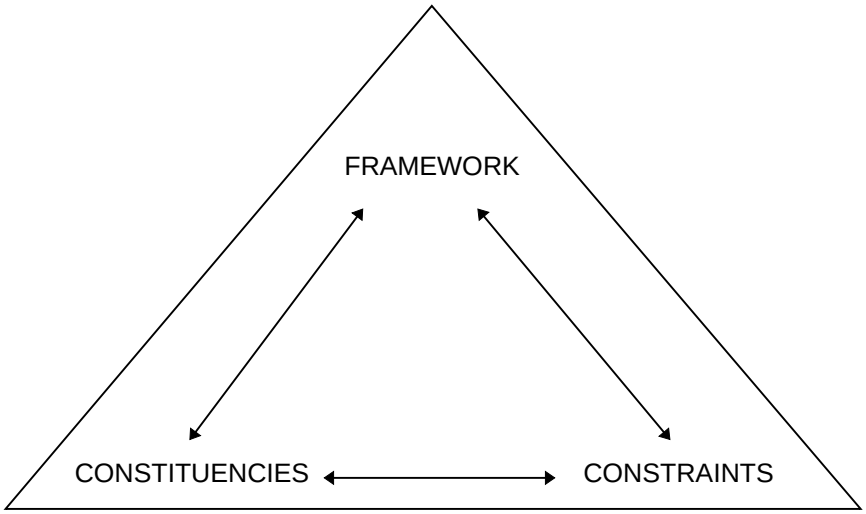

Figure 1. Components of the TOEFL 2000 project

First, it is important to consider the diverse constituencies served by the TOEFL program. College and university admissions officers who are choosing candidates for undergraduate and graduate programs need an expeditious way to screen the English-language proficiencies of large numbers of international students. Teachers of English as a second or foreign language (ESL/EFL), on the other hand, need more detailed, in-depth information to guide decisions about course placement and instructional design. Accordingly, they want TOEFL tasks to more closely approximate the types of communicative behaviors that are desired in university settings. Examinees use TOEFL test results as well to gauge their progress in instructional settings and to decide whether to apply to other programs. In developing an operational TOEFL 2000 test, then, there must be a clear understanding of the kinds of information that diverse test users need and want from the test.

Second, practical constraints need to be weighed. While exploring technological capabilities for constructing and delivering the next generation of a computer-based TOEFL test, for example, it is also necessary to balance the practical requirements involved in providing a secure test on demand to almost a million individuals each year.

Third, it is necessary to consider the framework for the test—the focus of this paper. By articulating a framework, we hope to provide a set of concepts and terms that participants in the TOEFL 2000 project can use to engage in future discussion and consensus-building. Our aim is to present the framework in enough detail to guide the development of more specific frameworks for discrete and integrative test tasks in the reading, writing, listening, and speaking modalities. Accordingly, this paper should be considered a work in progress. The ideas herein will evolve and be refined through subsequent research and continued discussions with TOEFL Policy Council and committee members, test development staff, researchers, colleagues in language teaching and testing, and others who use the TOEFL tests.

This paper consists of four sections: the background of the TOEFL 2000 project, the rationale for developing a framework, a description of work on the TOEFL 2000 framework to date, and plans for future work.

2. Background of TOEFL 2000 Project

In recent years, various constituencies,[2] including TOEFL committees and score users, have called for a new TOEFL test that: (1) is more reflective of communicative competence models; (2) includes more constructed-response tasks and direct measures of writing and speaking; (3) includes tasks that integrate the language modalities tested; and (4) provides more information than current TOEFL scores do about international students' ability to use English in an academic environment. Accordingly, in 1993, the TOEFL Policy Council[3] initiated the TOEFL 2000 project, a series of research and development activities that will lead to a new computer-based TOEFL test.

The impetus for this project comes from several sources. Many in the language teaching and testing communities associate the TOEFL test with discrete-point testing, which is based on the structuralist, behaviorist model of language learning and testing. As defined by Carroll (1961) and Oller (1979), discrete-point tests contain items that target only one element of a skill, such as grammar or vocabulary. ESL/EFL teachers are concerned that discrete-point test items, and the exclusive use of traditional, multiple-choice items to assess receptive skills, have a negative impact on instruction.

In addition, those who use TOEFL test scores in selecting students for undergraduate and graduate programs increasingly express concern that many international students who are admitted with high TOEFL test scores (i.e., above 550) arrive on campus with insufficient writing and oral communication skills to participate fully in academic programs. This underscores the need for direct measures of writing and speaking ability to be included in the TOEFL test.

Despite such concerns, the current TOEFL test continues to be widely viewed as the primary instrument for making admissions decisions. Other EFL tests—such as the University of Michigan's MELAB (Michigan English Language Assessment Battery) and the ELPT™ (English Language Proficiency Test) of the COLLEGE BOARD®—are, like the current TOEFL test, predominantly traditional, multiple-choice tests of receptive skills. Further, examinee volumes for these tests have been comparatively low to date, ranging from 2,500 to 8,000 annually.

---

[2] Committees directly involved in advising TOEFL 2000 include the TOEFL Committee of Examiners and the TOEFL Research Committee, two groups composed of second/foreign language testing and teaching experts. Constituencies primarily include score users from the North American college and university undergraduate and graduate admissions community, applied linguists, language testers, and second language teachers. While fewer in number, other users of TOEFL scores represent a diverse range of groups: public and private high schools, overseas colleges and universities, embassies, foundations and commissions, medical and professional boards, government agencies, language institutes, and a small number of private businesses.

[3] The TOEFL Policy Council is an independent, 15-member group that formulates policies governing the TOEFL program. Members of the Council represent the various TOEFL constituencies.

Although direct measures of speaking and writing abilities were incorporated into the Test of Spoken English (TSE) and Test of Written English (TWE®) during the 1980s, these tests are also used much less widely than the TOEFL test. For example, the average annual volume of TSE examinees is only 15,000, compared to more than 880,000 for the TOEFL test. The TWE test is only offered at 5 of the 12 TOEFL test administrations each year, and TOEFL score users report that this limited access to the test is the primary reason that they do not require TWE for admissions.

Several Canadian, British, and Australian EFL tests[4] do include some constructed-response tasks, but the annual testing volumes for these instruments are also quite low and the availability of testing centers and test administrations is relatively limited. Moreover, these tests have been criticized for their lack of evidence of scoring reliability and comparability of scores across test forms (Davies, 1987; Hamp-Lyons, 1987; Morrow, 1987; Rea, 1987; Tony, 1987; Alderson, Clapham, & Wall, 1995). Finally, there is little or no evidence that these tests are supported by an articulated theoretical model, or that they are more communicative or valid than the more traditional, multiple-choice tests.

Because the TOEFL test is so widely used in making admissions decisions, the impetus to revise the test has been strong. In recent years, certain aspects of the test were changed in an effort to address constituents' concerns. For example, single-statement listening comprehension items were eliminated from the test, the numbers of academic lectures and longer dialogs were increased, and vocabulary tasks were embedded in reading comprehension passages. Still, these changes reflected relatively minor progress toward an integrative approach to language testing.

More recently, in a more extensive effort to address the concerns enumerated earlier in this section, TOEFL program staff undertook several parallel, interrelated efforts with the advice and on-going review of the TOEFL Council and committees. Specifically, program staff systematically considered three broad areas: test users, technology relevant to test design and international delivery, and test constructs.[5] With respect to test users, TOEFL staff profiled examinees and score users, conducted a number of score user focus groups and surveys, and prepared reports on trends in international student admissions and intensive English program

[4] These include the Test in English for Educational Purposes (TEEP), Ontario Test of English as a Second Language (OTESL), Association of Recognised Language Schools Oral Examinations in Spoken English (ARELS), International English Language Testing System (IELTS), and University of Cambridge Local Examinations Syndicate (UCLES) certificate exams.

[5] Reports of efforts in these areas are being published in a monograph series. See Appendix A for a list of available monographs in each area.

enrollments.[6] These activities helped to clarify and elaborate on constituencies' concerns and needs. With respect to technology, staff examined existing and developing technologies in North American universities, as well as anticipated developments, that could facilitate implementation of computer-based testing (CBT).

The majority of the initial efforts, however, focused on test constructs and the development of prototype tests. Project staff systematically reviewed the literature on communicative competence and communicative language testing (Bachman, 1990; Bachman & Palmer, 1996; Canale, 1983, 1988; Canale & Swain, 1980; Cazden, 1996; Duran, Canale, Penfield, Stansfield, & Liskin-Gasparro, 1985; Ervin-Tripp, 1972; Halliday, 1970; Hymes, 1971, 1972a, 1972b; Munby, 1978; The New London Group, 1996; Savignon, 1983; Skehan, 1995; Stansfield, 1986; van Ek & Alexander, 1975; van Els & Engels, 1983; Weir, 1990; Wesche, 1987). The TOEFL Committee of Examiners (COE) elaborated on a model of communicative language use that was based on the communicative competence literature and committee discussions (Chapelle, Grabe, & Berns, 1997). These reports and expert opinions were intended to engage TOEFL committees and the larger language and measurement communities in considering the critical development and research issues for the TOEFL 2000 project. The framework proposed in this paper builds on these efforts.

After completing and reviewing commissioned papers on various constructs (Chapelle, Grabe, & Berns, 1997; Douglas, 1997; Ginther & Grant, 1996; Hamp-Lyons & Kroll, 1997; Hudson, 1996; Waters, 1996), and conducting further discussions with committees and project consultants, prototyping teams tried to operationalize critical aspects of communicative competence that were relevant to large-scale language testing. The teams created several modules of test tasks. The two primary purposes of the modules were to: (a) create sets of test tasks that took into account both the TOEFL Committee of Examiners' (COE) model and what TOEFL constituents had called for in a new TOEFL test, and (b) further identify and clarify research and development issues. The modules included extended reading and listening passages that were contextualized, were linked thematically, and contained integrated, performance-based writing and speaking tasks. As such, they represented a departure from the current TOEFL test in terms of content and types of tasks, the test development process, and scoring.

In 1995, the results of the cumulative efforts were reviewed (Taylor, Eignor, Schedl, & DeVincenzi, 1995) and discussed with TOEFL committees and project consultants. It became apparent that, while the COE model and other construct papers expanded and elaborated on existing models of communicative competence, they did not provide a framework that test developers could use to construct forms of a new test. Others have also found it difficult to connect the existing research on communicative competence to a test framework. McNamara

[6] While test user profiles and market survey data are considered proprietary to the TOEFL program, enrollment trend information is being released in a TOEFL monograph (Powell, in press).

(1996) provides an extensive critique of attempts to build tests using communicative competence models and underscores the difficulty of linking theory and testing practice.

The review of TOEFL 2000 efforts also indicated that there was no clear model for providing the kinds of test information that score users were requesting. Moreover, many critical psychometric and technical issues remained unresolved. For example, should existing psychometric models be expanded or should new models be developed to deal with test tasks that integrate two or more modalities? How should score reporting be enhanced, by applying existing procedures or by developing new ones? To support computer administration of the test, how can the process of developing large item pools be made more efficient? Thus, it became apparent that a longer development and implementation time frame would be required to create, validate, and implement a new TOEFL test that responds to constituencies' concerns and needs.

The resulting consensus was that the development strategy should be split by moving the current TOEFL test with some important design enhancements to CBT in 1998 while continuing to pursue the original vision of TOEFL 2000 within the Research Division of ETS.

The decision to introduce an interim computer-based test provided an opportunity for the TOEFL test to benefit from the foundational TOEFL 2000 research and development efforts. New test design features were identified (ETS, 1994, 1996), computer-based prototypes were created and piloted, a study of TOEFL examinees' computer familiarity and performance on computer-based test tasks was undertaken (Kirsch, Jamieson, Taylor, & Eignor, 1998; Eignor, Taylor, Kirsch, & Jamieson, 1998; Taylor, Jamieson, Eignor, & Kirsch, 1998), and TOEFL CBT implementation plans were developed.

Correspondingly, by creating a computer-based infrastructure and platform, TOEFL CBT will enable program staff to explore future innovations in test design for TOEFL 2000. Hence, as work on developing and implementing the computer-based TOEFL test proceeds, research resources are being directly applied to building the essential assessment framework for TOEFL 2000, the next generation of computer-based TOEFL tests.

The remaining sections of this paper lay out the rationale for and conceptualization of the TOEFL 2000 framework.

3. Conceptualizing a Framework

Instruments like the TOEFL test typically rank-order examinees by assigning them numerical values (i.e., test scores) without identifying the proficiencies that underlie the numbers. In other words, the test scores tell us that examinees differ without telling us *how* they differ. The general purpose of developing a framework for the TOEFL 2000 test is to improve the measurement of second-language competence. Rather than merely assign numerical values or positions to examinees based on their responses to a set of tasks, the goal is to give meaning and interpretability to the numbers (Messick, 1989).

In considering the development of the TOEFL 2000 framework, we identified what we consider to be a set of necessary components:

- A framework should begin with a general statement of the test's purpose, one that guides what the test will measure and how scores should be used.

- A framework should identify various task characteristics and indicate how these characteristics will be used in constructing language tasks.

- Variables associated with each task characteristic should be specified, and research must be conducted to show that these variables account for large percentages of the variance in test performance. Variables that appear to have the largest impact on test scores should be used to create an interpretive scheme. This is a crucial step in the process of measurement and validation.

While the chief benefit of constructing and validating a framework for the TOEFL 2000 test is improved measurement, a number of other potential benefits are also evident. Namely:

- A framework provides a common language and a vehicle for discussing the purpose of the test and what it is trying to measure.

- Such a discussion allows us to build consensus around the new framework and measurement goals.

- A framework sets the boundaries and parameters for constructing language tasks and for interpreting scores.

- Identifying and understanding particular variables that underlie successful performance on a set of language tasks further our ability to construct new tasks that more fully represent the domain(s) being assessed.

- Identifying particular variables that are related to task performance makes it possible to train consultants to develop specific types of assessment tasks. Such tasks should require less refinement by ETS test developers, thereby reducing the costs of constructing new tasks.

- Knowing and understanding the variables, and how they contribute to successful performance, provide an overall scheme for interpreting scores along a proficiency scale or domain. This enhances the construct validity of the test and moves us closer to the type of measurement referred to by Messick.

- An analysis of the kinds of knowledge and skills associated with successful performance provides a basis for establishing standards or levels of proficiency. As we increase our understanding of what is being measured and our ability to interpret scores along a particular scale, we have an empirical basis for communicating a richer body of information to examinees, admissions officers, language teachers, and other test users.

- Linking research, testing, practice, and public policy promotes not only the continued development and use of the test, but also understanding of what it is measuring.

There are many ways to develop a framework, but we believe it is essential to follow a process of identifying and validating critical features that can be used to predict successful performance. These can then be used to create an interpretative scheme, develop task specifications, build and evaluate prototype tasks, and ultimately produce test specifications.

Any framework needs to be validated, and this requires both qualitative and quantitative studies. The TOEFL 2000 framework must be discussed and revised in the TOEFL community, and focus groups should be conducted to determine the kinds of information that score users want and the ways in which that information should be communicated.

4. The Framework: Work to Date

The framework for the TOEFL 2000 test consists of six parts, shown in Figure 2. These parts represent a logical sequence of steps that must be addressed, from wanting to build a test, to having specifications for a new test, to providing an empirically based interpretation of test scores.

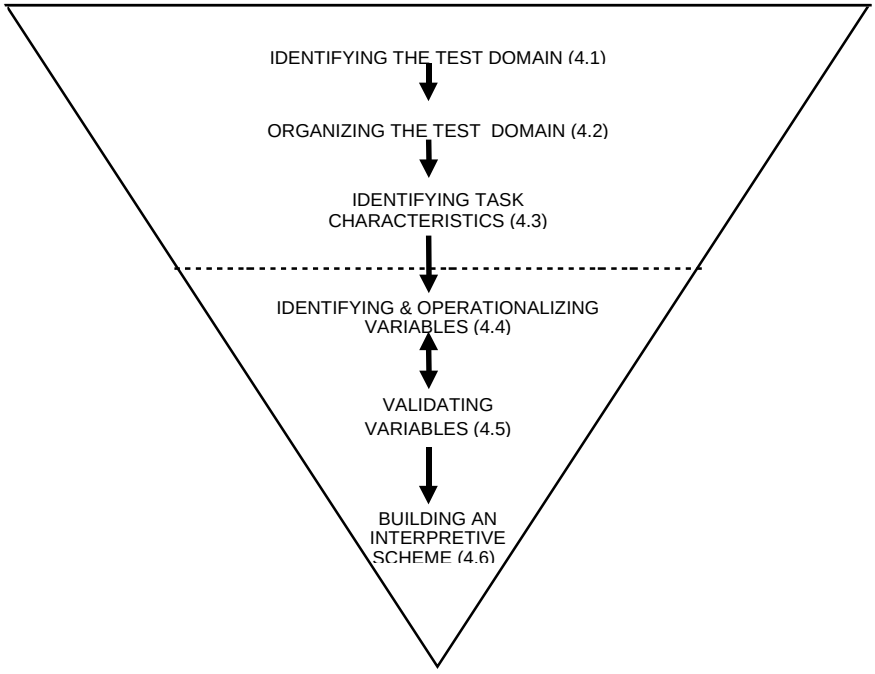

Figure 2. The structure of the TOEFL 2000 framework

The first three components of the framework—identifying and organizing the domain, and identifying task characteristics—are discussed in the following sections. Part 4.1 defines the assumptions underlying the test and describes its purpose. Part 4.2 provides an organizational scheme for second language communicative ability, the domain of interest in TOEFL 2000. Part 4.3 identifies key task characteristics based on a review of literature on communicative competence and on foundational work for the TOEFL 2000 project. These three components of the framework are considered to be set at this point. They were developed based on discussions with TOEFL program staff and constituencies and on our review of the literature and previous TOEFL 2000 efforts.

The second three components of the framework—identifying and operationalizing variables, validating variables, and building an interpretive scheme—stem from relevant reading and adult literacy research and from further consideration of the construct-related TOEFL 2000

monographs. These components are not considered set at this point. Part 4.4 discusses how the validation process could be conducted, describing how to begin identifying and operationalizing variables associated with the various task characteristics outlined in part 4.3. Part 4.5 sets out a procedure for validating the variables and assessing the contribution of each variable in the measurement model. The final part, 4.6, discusses how an interpretive scheme can be built from the variables. As the project moves forward, specific variables may be revised and refined, and other methodological approaches to validation may be applied.

4.1 Identifying the Test Domain

Internal staff and external consultants have held a number of working meetings to review the purpose of the TOEFL test as it relates to the development of a TOEFL 2000 framework. Several assumptions have emerged from these meetings, and these have led to a statement of purpose which identifies the domain of the TOEFL 2000 test:

> *The purpose of the TOEFL 2000 test will be to measure the communicative language ability of people whose first language is not English. It will measure examinees' English-language proficiency in situations and tasks reflective of university life in North America. Abilities will be reported along one or more scales characterized by increasing levels of proficiency. Scale scores are designed to be used as one criterion in decision making for undergraduate and graduate admissions. Information derived from the proficiency levels may also be used in guiding English-language instruction, placement decisions, and awarding of certification.*

This statement of purpose provides the basis for creating the framework to be used in test development. Accordingly, it is important to consider each part of the statement in turn.

The purpose of the TOEFL 2000 test will be to measure the communicative language ability of people whose first language is not English.

The TOEFL 2000 framework needs to be anchored in theories of communicative competence. Hymes, who is generally credited with introducing the concept of communicative competence, did so to identify language abilities that are acquired through training and practice. This contrasts with Chomsky's notion of grammatical competence as an inherent property of individuals. In identifying competence as "the most general term for the capabilities of a person," Hymes (1972b, p. 182) suggests that his notion of competence is closely related to the notion of ability in the psychological literature. As Messick (1981) points out, ability refers to a personal attribute that "by virtue of its implicit transfer potential appears more oriented to the future" (p.16). Competence, in contrast, "seems anchored in the present, referring to something one does and can do" (Messick, 1981, p. 16).

The goal of the TOEFL 2000 test must be to provide information about future performance, about ability. Hence, although the literature we borrow from refers to communicative

competence, we believe that communicative language ability is a more appropriate term for the attribute we want to measure. Thus, we arrive at the same position as Bachman (1990), though by a much different route.

The TOEFL 2000 test will measure examinees' English-language proficiency in situations and tasks reflective of university life in North America.

This means that the test will focus on the application of language skills in selected situations rather than on isolated skills. As a result, the tasks designed for the TOEFL 2000 test will simulate the range of English-language proficiency expected of students in North American universities. One consequence of this decision is that the TOEFL test may not continue to include a separate measure of structure. Instead, the knowledge and skills associated with this aspect of the test will most likely be embedded in tasks that are developed around the various modalities.

Abilities will be reported along one or more scales characterized by increasing levels of proficiency.

There appears to be a strong desire on the part of institutional score users to have scores reported in each of four areas: reading, writing, speaking, and listening. The models we use should enable the development of language tasks that measure these areas both independently and integratively. Thus, for example, some reading tasks should measure various aspects of processing printed information, while others should require examinees to read something and then write about or tell someone about what they have read.

A major challenge in developing a new TOEFL test will be to identify a set of variables that account for a significant amount of variance in the distribution of task difficulty along each scale. The understanding that comes from this process provides a model that can be used to explain increasing task difficulty, interpret scores, and empirically determine levels along each scale that mark shifts in examinees' ability to perform various kinds of tasks.

Scale scores are designed to be used as one criterion in decision-making for undergraduate and graduate admissions.

This is in keeping with the current purpose of the TOEFL test. Proficiency scales provide more information on which to base admissions decisions, however. Admissions officers and committees will receive both norm-referenced and criterion-referenced information. That is, the scores will be interpretable both in terms of how well individuals perform compared to other examinees and in terms of scales that have been anchored by tasks and proficiency descriptions.

Information derived from the proficiency levels may also be used in guiding English-language instruction, placement decisions, and awarding of certification.

Although the TOEFL test will continue to be used primarily to guide decisions about admissions to undergraduate and graduate programs of study, future research based on the new proficiency scales may provide information that will expand the valid uses of the scores and thus serve to broaden the purpose of the test.

4.2 Organizing the Test Domain

Having identified the domain of communicative language ability in the statement of purpose, it is necessary to decide how to organize the domain. This organization will affect test design and development as well as score reporting. In short, it will serve as the frame of reference for the TOEFL 2000 test.

Different approaches to organizing language can be taken. The most traditional way is by modality: reading, listening, speaking, and writing, and perhaps also grammar and vocabulary (e.g., Handschin, 1923; Coleman, 1934; Long & Richards, 1987; Grosse, 1991). Alternatively, one could use a theoretical perspective, categorizing language in terms of grammatical, discourse, sociolinguistic, and strategic competence (Canale & Swain, 1980; Canale, 1983; Duran et al., 1985). Or, one could organize language in terms of organizational competence (grammatical and discourse) and pragmatic competence (illocutionary and sociolinguistic) (Bachman, 1990; Hudson, Detmer, & Brown, 1995). Still another option is to organize language by setting and communicative tasks (e.g., Wesche, 1987; Pica, Kanagy, & Falodun, 1993; Bachman & Palmer, 1996; Skehan, 1995; McNamara, 1996; Chapelle et al., 1997).

Based on a review of the various approaches, it was decided that the most practical way to organize language tasks for the TOEFL 2000 framework is by modality. Thus, the test will include measures of speaking, writing, listening, and reading. Within these four areas, the new test will include a variety of language features, including not only grammar and vocabulary but also discourse, pragmatics, and sociolinguistics, as well as setting and task.

There were several reasons for choosing modality as the organizing principle for the test. In surveys of score users such as admissions officers and graduate deans, respondents requested that scores be reported for speaking, writing, listening, and reading because these skills relate to the kinds of decisions they need to make (Taylor, 1993; Ginther & Grant, 1996). While there is movement toward content- or task-based ESL/EFL curricula, the four modalities are still strongly represented in the preparation of ESL teachers, as well as in the curricula used in intensive English-language programs (Grosse, 1991; Brown, 1996).

It is important to note that speaking, writing, listening, and reading can be tested both integratively and independently. At this time, it is envisioned that a number of the TOEFL 2000 tasks will assess the skills integratively, using combinations such as reading a text and writing a summary, listening to a question and providing a spoken response, or reading an article, listening to a lecture, and comparing and contrasting information in an essay. These integrated tasks will provide information about examinees' ability in more than one skill area. The TOEFL 2000 test will also assess reading, writing, listening, and speaking skills independently. We

expect that information from the integrated tasks will be combined with information from the independent tasks to construct a profile of language abilities for each examinee.

4.3 Identifying Task Characteristics

Thus far, we have described the domain of the TOEFL 2000 test with respect to the four modalities: reading, writing, speaking, and listening. The next step is decide which task characteristics to include in the test framework.

We begin with Bachman and Palmer's (1996) notion that a finite number of task characteristics influence students' communicative competence in a given language, as measured by a test. These characteristics, shown in Figure 3, include situation, text material, and test rubric.

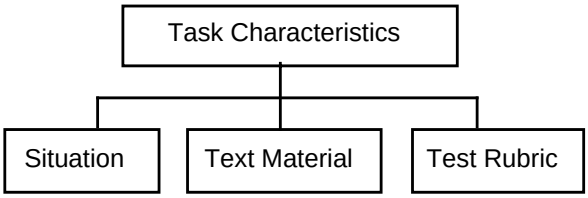

Figure 3. A model of task characteristics

Situation is the "extralinguistic setting in which an utterance takes place" (Crystal, 1991, p. 318). Most applied linguists today agree that a description of communicative language ability must include language use in context (Hymes, 1972a, 1972b, 1974, 1996; Canale & Swain, 1980; Bachman, 1990; Bachman & Palmer, 1996; Chapelle et al., 1997).

Text material refers not only to the language that the examinee reads in a reading task or listens to in a listening exercise, but also to the text produced by an examinee when writing or speaking. While no one would doubt that a language test should include language material, what is critical to design and interpretation are the specific features of the text material that are taken into account in constructing test tasks. These are the text material variables.

Test rubric refers to characteristics of the questions or directives that set out the language tasks for examinees, the response formats, and the rules for scoring examinees' productions. Generally, the questions and directives will specify a purpose or goal and the material to be used in reaching that goal. The TOEFL 2000 test will not rely solely on multiple-choice tasks but will include open-ended and constructed-response tasks.

Once task characteristics have been identified for the test, it is necessary to further define and operationalize these characteristics—the objective of the next section.

4.4 Identifying and Operationalizing the Variables

In operationalizing the task characteristics of situation, text materials, and test rubric, we further delimit the domain, for two reasons. First, identifying variables for each of the task characteristics will provide a set of criteria that can be used to evaluate existing tests. Second, and more importantly, these variables provide a foundation for building new tests and for interpreting scores on those tests.

Readers should keep in mind that the variables discussed in this section represent an initial attempt on our part. We anticipate a process in which teams will be formed for each of the four modalities and charged with extending, refining, and perhaps changing the variables presented here, based on participants' expert judgments and empirical verification.

4.4.1 Situation

Situation encompasses the extralinguistic elements associated with language tasks (cf., Crystal, 1991, 1992). As shown in Figure 4, in operationalizing this task characteristic, we include the following five variables: participants, content, setting, purpose, and register.

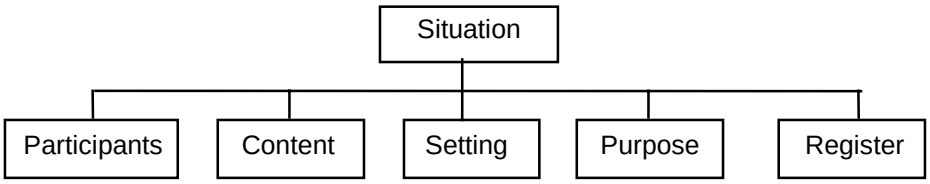

Figure 4. A model of task characteristics: situation

Participants

The participants are the people involved in the language act and the relationships among them. The participants variable can be operationalized in terms of gender, ethnicity, age, and role. This allows for language situations involving males and females of varying ages and in various roles (for example, students, faculty, and staff).

Content

Content refers to the subject matter included in language tasks. This subject matter must be general enough to ensure that discipline-specific knowledge is not the primary factor affecting performance. In addition, it must be broad enough to encompass a range of topics and language. Based on these constraints, we suggest that the TOEFL 2000 test include three basic types of subject matter: academic content, class-related content, and extracurricular content.

The academic content of the test should be restricted to the natural and social sciences because these disciplines cover a wide range of genres and yet constitute fields of general knowledge. Class-related content is defined in terms of the everyday demands that students encounter as they juggle their university classes and assignments. Tasks may include identifying the location of places or materials or using class or meeting schedules. Extracurricular content consists of topics that frequently arise for students but are not related to their classes, such as illness, the weather, or roommate relations.

Setting

Setting is defined as the place where the language act occurs. In order to meet the purpose of the TOEFL test and yet be able to sample a broad range of language, we propose that three types of setting be represented: instructional milieu, academic milieu, and non-academic milieu. (These locations are based on a description of academic and social use contexts developed by Duran et al., 1985.)

The instructional milieu includes all places where formal instruction takes place, such as lecture halls, labs, seminar rooms, and classrooms. Academic milieu is intended to indicate typical places outside of the classroom where aspects of academic life are dealt with, such as a study room in a dormitory, the library, an instructor's office, the bookstore, a writing center, or a computer center. Non-academic milieu includes places that are not usually associated with academic content, but where social and business transactions take place. This category could include the business office, international students' office, and the health center, as well as dormitory rooms and dining areas.

Purpose

Purpose is defined as the reason why we engage in tasks. Language purposes have been variously categorized by Bachman (1990), Heath (1980), and the Council of Europe (van Ek & Alexander, 1975), but these are seen as variations on Halliday's (1973) list. For the TOEFL 2000 test, six of Halliday's seven categories relate to the purposes for which international students would use English in a North American university: heuristic, instrumental, regulatory, personal, representational, and interactional. (See Appendix B for a description and example of each.) Halliday's seventh category, imaginative, is considered beyond the scope of the TOEFL 2000 test.

Register

Based on the work of others in the field (e. g., Tarone, 1983; Halliday & Hassan, 1976; Douglas, 1997, personal communication), register is defined here as the degree of formality that is used in language. Three degrees of formality seem applicable to the TOEFL 2000 test: formal, consultative, and informal. We can think of these as forming a continuum.

A formal register is used in textbooks, formal lectures, prepared class presentations, and term papers. There is little or no observable real-time interaction between the speaker and listener or writer and reader. Shared background knowledge is assumed at an academic/professional level, not a personal one. Consultative registers are used in official business such as in an office, in business letters, or over the phone with strangers. Here, a considerable amount of background information has to be provided. There is interaction between co-participants, although turns may be quite lengthy. An informal register is used socially and can be emotional. The highest degree of interaction is present, and turns are often short and elliptical, as a large amount of background information is assumed to be shared.

Summary

Each of the five situation variables—participants, content, setting, purpose, and register—represents an important aspect of language use, and all of these variables are simultaneously at play in language tasks. As one example, consider a professor who is delivering a lecture on amber to a botany class. The setting is instructional milieu; the participant is a middle-aged male professor; the register is formal; the purpose is heuristic; the content is natural science. If, after class, a student were to visit the professor in his office and ask for advice on where to find resources for a required term paper, a different situation would arise. The setting would be academic milieu; the participants would be a 20-year-old female student and the professor; the register would be consultative; the purpose would be instrumental; the content, finding references, would be class-related. By including situation as one task characteristic within the TOEFL 2000 framework, we hope to include contextualized language use and to systematically evaluate how this aspect of a language task contributes to task difficulty.

4.4.2 Text Material

As shown in Figure 5, the text material—that is, texts for reading, scripts for listening, prompts for speaking and writing—can be described by three types of features. Grammatical features relate to the structure of the sentences and the vocabulary used in the text. Pragmatic features relate to the intent of the text's creator. Discourse features relate to the nature and structure of the text as a whole, including rhetorical type and textual organization.

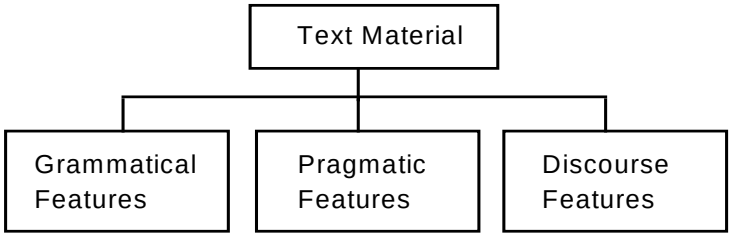

Figure 5. A model of task characteristics: text material

Grammatical features

Numerous grammatical features of a text[7] influence the difficulty of tasks based on that text. These features involve both syntax and vocabulary.

In deciding which grammatical features to include as variables, one requirement is that they must be likely to have an impact on variety and difficulty. A second requirement is that it must be easy to code the features, preferably by computer. Examples of possible grammatical variables include distribution of sentence types, word classes and verb forms, types of subordinate clauses, readability scores, and amount of information.[8] Studies of the distribution of these syntactic features in a variety of discourse settings are available and should be used in setting the range standards for TOEFL 2000 (see, for example, the papers in Selinker, Tarone, & Hanzeli, 1981).[9] These variables are explained in more detail in Appendix C. Note that the range of values to be included for these variables (or others) will differ from skill to skill, and therefore must be determined by the modality teams referred to earlier.

It is also desirable to monitor a number of vocabulary features. These, too, must be features that can be easily coded, ideally by a computer. The distribution of words by frequency should reflect the frequency in university-level texts. The distributions developed from the Brown corpus (Francis, Kucera, & Mackie, 1982; Kucera & Francis, 1967) are appropriate for comparisons.[10] Also of interest are the proportion of abstract to concrete words and the proportion of common to technical words (see Appendix C).

Pragmatic features

Most of the texts included in the TOEFL 2000 test will be expository in a broad sense (see the discussion of purpose in the situation specifications above), so the dominant pragmatic function will be to impart or seek factual information (exposition) or to present or defend an analysis (argument). A small number of expository texts may have the expression or defense of

[7] Unless otherwise indicated, "text" refers to both written and spoken material. When computer-coding of oral texts is mentioned, it is assumed that the coding will be performed on a written transcript of the material.

[8] Studies suggest that the syntactic features of text have little impact on difficulty when they are considered as features of the text alone. They contribute significantly to task difficulty, however, when they are viewed in connection with the characteristics of directives for the tasks to be performed using that text (Freedle & Kostin, 1993).

[9] As these studies have not been consolidated in a useful way, one productive TOEFL 2000 project would be a review of these studies and their relevance for test design and construction.

[10] No "off-the-shelf" programs currently exist to do this, but the Natural Language Processing group at ETS believes it would not be difficult to adapt existing software to produce the desired counts.

attitudes as their main function. While most texts will be persuasive to some extent, few will be exclusively so. A small number of texts, particularly those selected for non-academic milieu situations, will have socialization as their primary function and may be in the form of a narrative or dialogue. Thus, the primary intent of the author of the text material can be categorized as expository, argumentative, persuasive, or socializing.

Discourse features

Two kinds of discourse properties need to be identified for each text in the TOEFL 2000 test. Rhetorical properties are patterns or formulas related to the author's goal in producing the text (or part of the text). For the most part, these are common rhetorical classifications. Text structure refers to mechanisms for relating parts of the text to each other. Structure is often closely related to the question or directive; a certain text structure may limit the sorts of questions or directives that can be created for that text.

Rhetorical properties.[11] Written and oral texts can be generally classified as definition, description, classification, illustration, cause/effect, problem/solution, comparison/contrast, regulatory, or analysis (see Appendix D). Although there is no definitive list of rhetorical types, this set seems reasonably representative.[12] Most professional texts have both major and minor rhetorical properties. Thus, a text that is primarily a description of a mechanism may also include some definitions or a short description of a process. As Hale, Taylor, Bridgeman, Carson, Kroll, and Kantor (1996) have shown, it is possible to classify writing assignments using these types of properties, and it is reasonable to assume that the longer productions required in a speaking test can also be so classified. It is as important to apply these specifications to productive tasks as to receptive ones.

Short oral texts, especially those involving interaction, are not well classified with these categories. Rather, interactions are defined by rules for turn-taking by participants (Sacks, Schegloff, & Jefferson, 1974). Although no definitive list exists, some common types of turns—called adjacency pairs in the conversational analysis literature (Schegloff & Sacks, 1973)—include assessment-concurrence, invitation-acceptance, and compliment-downgrade. Appendix D, Table D2, offers one example of what might be developed from the literature. It would be worthwhile to develop a more comprehensive set as the rules for adjacency pairs are related to the rules for conversational politeness, which differ from culture to culture (Brown & Levinson, 1978).

[11] Hale et al. (1996) refer to these as patterns of exposition.

[12] We have found technical writing texts to be most useful in explicating these types. In particular, this discussion borrows from Lannon (1982); Mills and Walter (1978); and Pickett and Laster (1975).

Text structure properties. These properties concern the ways in which information in the text is related. It is necessary to distinguish three types of text: documents, prose, and interactions. Text features that display relationships can be either typographical, such as headings and table layout, or syntactic, as when they are signaled by explicit discourse markers. In general, documents are dominated by typographical features, and prose texts, by syntactic features. These types of texts are introduced here. (For a more detailed explanation, see Appendix E.) It should be noted that, as with rhetorical properties, texts can be a mixture of these three types of structure. Most textbooks, for example, are primarily prose texts, but many include document parts, such as tables, graphs, and other illustrations. A lecture (an oral prose text) may have interludes of interaction, as when the lecturer responds to questions.

a. Documents

Documents are written texts[13] that consist of words, phrases, and/or diagrams and pictures organized typographically. The most useful categorization of document structures is the list model developed by Kirsch and Mosenthal.[14] According to this model, the basis of a document is a list of elements that have some organizing category in common. More complex documents are constructed by combining lists in several ways. Briefly, documents are described as types of lists, ranging from simple to combined, intersecting, nested, and combination. Tables, schematic diagrams, schedules, and graphs are typical academic documents that can be categorized as one of these list types. (See Appendix E, Tables E1 and E2, for definitions and examples.)

b. Prose

Prose texts are written or oral texts that consist of sentences organized into paragraphs (or their oral counterpart). Typical examples in academic settings are lectures and textbooks. The prose texts in the TOEFL 2000 test will be primarily expository (see pragmatics discussion in this section and purpose discussion in the section on setting). These texts present information that readers or listeners can use to improve their knowledge of a topic. Mosenthal (1985) has presented the most complete model of expository prose structure. In particular, we are concerned with distinguishing more loosely structured texts from more tightly structured ones. It is also important to distinguish texts that contain explicit signals (headings and sub-headings) from those that do not.

[13] Documents do not normally occur as oral texts. An exception would be when a written document is read aloud, as when an instructor lists marriage types in an anthropology class.

[14] The Kirsch and Mosenthal model was set out in detail in a series of articles in the *Journal of Reading* between 1990 and 1992. See References for the complete list.

Like the model of document structure discussed earlier, Mosenthal's classification of prose text structure is based on relations expressed in the text. The types, which are arrayed from the lowest to the greatest amount of in-text structure, include records/reports, generalized reports/records, loose classifications, strong classifications, speculatives, and theoreticals (see Appendix E, Table E3). These structures are hierarchical. A theoretical text almost certainly includes speculative material, which in turn requires at least a loose classification. Most university lectures, particularly those designed for introductory courses, are loose or strong classifications. At higher levels (last undergraduate years and graduate programs), lectures increasingly incorporate speculative and theoretical texts.

c. Interactions

Interactions are oral texts that are organized by the rules of turn-taking (Sacks et al., 1974). Actual conversations typically consist of sentence fragments together with full sentences, while in test settings, complete sentences are almost always used. Three elements are important in the structure of interactions: turn-taking, topic, and function (see Appendix E, Table E4). These are not classifications, as are the types of lists, but are components of every interaction.

4.4.3 Test Rubric

Test rubric is the third task characteristic that will be examined as part of the TOEFL 2000 test (Figure 6). It includes three sets of variables: questions and directives, response formats, and rules for scoring responses.

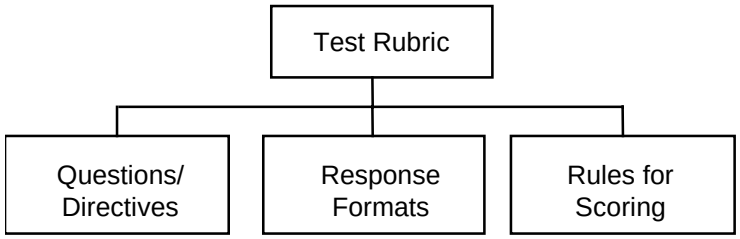

Figure 6. A model of task characteristics: test rubric

Various classifications of test questions and directives have been proposed in the education and testing literatures (Bloom, 1956; Pearson & Johnson, 1978). Typically, these classification systems have been developed independent of text characteristics and tend to have low correlations with task difficulty because they fail to take text features into account. For example, one question about "main idea" could be very easy because it simply requires the examinee to locate a statement in an opening paragraph. On the other hand, another main idea question might be very difficult because it requires the examinee to draw inferences from several paragraphs of

information. More recently, second-language testers have recognized the importance of the relationship between the wording of test questions and text features, but they have not proposed ways to quantify these interactions (Bachman & Palmer, 1996).

In this section of the paper, we discuss research[15] that has attempted to quantify relationships between test questions and text features that account for large percentages of variance within a set of language tasks. This research is intended to provide a starting point for further discussion by the TOEFL 2000 modality teams.

Questions and Directives

In the following pages, we explore relationships between test questions and directives (hereafter, for the sake of brevity, both are referred to simply as "questions") and text. We describe these relationships in terms of three continua—type of information requested, type of match, and plausibility of distractors—which are drawn from reading/literacy research and have been applied to some of the experimental reading and listening tasks developed for the computer–based TOEFL test. Although the three continua might also capture the information-processing properties of speaking and writing tasks, we have not yet investigated this. Therefore, the ensuing description focuses primarily on reading tasks, with some references to listening tasks.

Type of Information Requested refers to the kinds of information that readers and listeners must identify to answer a test question successfully. The more concrete the requested information, the easier the task is judged to be. In previous research based on large-scale assessments of adults' and children's literacy (Kirsch, 1995; Kirsch & Mosenthal, 1995; Kirsch, Jungeblut, & Mosenthal, in press), the type of information variable was scored on a 5-point scale. A score of 1 represented information that was the most concrete and therefore the easiest to process, while a score of 5 represented information that was the most abstract and therefore the most difficult to process. For instance, questions that asked examinees to identify a person, animal, or thing (i.e., imaginable nouns) were said to request highly concrete information and were assigned a value of 1. Questions asking respondents to identify goals, conditions, or purposes were said to request more abstract types of information. Such tasks were judged to be more difficult and received a value of 3. Questions that required examinees to identify an "equivalent" were judged to be the most abstract and were assigned a value of 5. In such cases, the equivalent tended to be an unfamiliar term or phrase for which respondents had to infer a definition, interpretation, or predicating condition from the text. Appendix F provides a more detailed description of scoring for the type of information variable.

Type of Match refers to the way in which examinees process text to respond correctly to a question. It includes the processes used to relate information in the question to the necessary

[15] See References for the complete list of relevant research conducted by Kirsch, Mosenthal, and Kirsch et al.

information in the text as well as the processes needed to either identify or construct the correct response from the information available.

Four types of matching strategies were identified: locating, cycling, integrating, and generating. Locating tasks require examinees to match one or more features of information stated in the question to either identical or synonymous information provided in the text. Cycling tasks also require examinees to match one or more features of information, but unlike locating tasks, they require respondents to engage in a series of feature matches to satisfy conditions stated in the question. Integrating tasks require examinees to pull together two or more pieces of information from the text according to some type of specified relation. For example, this relation might call for examinees to identify similarities (i.e., make a comparison), differences (i.e., contrast), degree (i.e., smaller or larger), or cause-and-effect relations. This information may be located within a single paragraph or it may appear in different paragraphs or sections of the text. In integrating information, examinees draw upon information categories provided in a question to locate the corresponding information in the text. They then relate the text information associated with these different categories based upon the relation term specified in the question. In some cases, however, examinees must generate these categories and/or relations before integrating the information stated in the text.

In addition to requiring examinees to apply one of these four strategies, type of match between a question and the text is influenced by several other processing conditions that contribute to a task's overall difficulty. The first of these is the number of phrases that must be used in the search. Task difficulty increases with the amount of information in the question for which the examinee must search in the text. For instance, questions that consist of only one independent clause tend to be easier, on average, than those that contain several independent or dependent clauses. Difficulty also increases with the number of responses that examinees are asked to provide. Questions that request a single answer are easier than those that require three or more answers. Further, questions that specify the number of responses tend to be easier than those that do not. For example, a question that states, "List the 3 reasons . . ." would be easier than one that said, "List the reasons . . . ." Tasks are also influenced by the degree to which examinees have to make inferences to: a) match the given information in a question to corresponding information in the text, and b) identify the requested information. A more detailed explanation and a flow diagram for scoring the type of match variable are provided in Appendix G.

Plausibility of Distractors concerns the extent to which information in the text shares one or more features with the information requested in the question but does not fully satisfy what has been requested. Tasks are judged to be easiest when no distractor information is present in the text. They tend to become more difficult as the number of distractors increases, as the distractors share more features with the correct response, and as the distractors appear in closer proximity to the correct response. For instance, tasks tend to be judged more difficult when one or more distractors meet some but not all of the conditions specified in the question and appear in a paragraph or section of text other than the one containing the correct answer. Tasks are judged to be most difficult when two or more distractors share most of the features with the correct

response and appear in the same paragraph or node of information as the correct response. A more detailed description of these continua is provided in Appendix H.

Response Formats

The testing literature provides little guidance on how different response formats—multiple-choice vs. constructed response, for example—actually affect examinees' performance on a test. Traub (1993) reviewed nine studies of the differences between multiple-choice and constructed-response test items. Although some of the studies found differences between the formats, the differences were small,[16] especially in the studies of reading comprehension. In summing up the research on reading tests, Traub concluded that "the answer is that [reading comprehension] tests that differ by format do *not* measure different characteristics" (1993, p. 38, italics original). More importantly for construct validity, Traub found that none of the studies that found a difference could identify the construct differences between the two formats.

Traditionally, the TOEFL program has used only multiple-choice items, although the TWE and TSE test include various kinds of constructed-response tasks. While the TOEFL CBT sought to expand the range of select-type items beyond the usual four-option multiple-choice format, no CBT tasks, except for the writing prompt, required a constructed response. The questions for the TOEFL 2000 test, then, are:

- Do new response formats provide better measurement of the existing constructs?
- Can new response formats be developed that will allow measurement of new constructs, such as integration across modalities?
- What is the construct difference, if any, between different response formats?

The introduction of new response types through the computer-based TOEFL platform will provide one stream of information for TOEFL 2000 design, but additional supporting research directed at these three questions will be required. We would expect the impact of any new response type on examinee performance to be thoroughly examined before its incorporation into the framework.

Rules for Scoring

The TOEFL 2000 modality teams will have to address numerous scoring issues. The choice of scoring format and procedures can have an effect on the test's ability to measure different

[16] The greatest differences were found in studies that compared multiple-choice writing tests, such as the Test of Standard Written English (TSWE), with essay tests.

constructs. For example, primary trait and holistic scoring of compositions may yield information about different constructs even when used with the same writing prompts. It will be important to understand the kinds of score interpretations supported by different procedures and to select the appropriate procedure(s) in light of these interpretations.

Similarly, dichotomous and partial credit scoring may provide different distributions of task difficulties and examinee scores. In this case, as well, it will be necessary to document the relationship between the scoring model and the interpretation of performance that is needed for the TOEFL 2000 test and to select the procedure that best supports the interpretations required.

Finally, it should be noted that some examinees might adapt their performance to fit a particular scoring scheme if they know about that scheme in advance. Since at least some examinees are likely to know about the scoring procedures, it will be necessary to implement some process to ensure that all examinees are informed about the scoring system to ensure equity.

4.5 Validating the Variables

In the previous section (4.4), we described our initial efforts to operationalize the task characteristics of situation, text materials, and test rubric, and identified possible variables for each of these task characteristics. This part of the framework describes a process for validating the variables. While researchers (Duran et al., 1985; Bachman & Palmer, 1996) have identified a range of variables that may contribute to adults' communicative language ability, what has been generally lacking in these efforts has been the specification and validation of variables that define the communicative competence components of either language tasks or specified levels of ability.

In the related area of adult literacy, however, six studies have reported on adults' communicative language ability in the domain of reading:

- The Young Adult Literacy Assessment (Kirsch & Jungeblut, 1986),
- The U.S. Department of Labor's Literacy Assessment (Kirsch & Jungeblut, 1992),
- The ETS Tests of Applied Literacy Skills (ETS, 1991),
- The U.S. Department of Labor's Workplace Literacy Tests (Kirsch, Jungeblut, & Campbell, 1991),
- The National Adult Literacy Survey (Kirsch, Jungeblut, Jenkins, & Kolstad, 1993), and
- The International Adult Literacy Survey (Kirsch, 1995).

Reading tasks for these large-scale assessments were developed to represent the broad range of purposes for which adults read expository and narrative materials in occupational, social, and

educational settings (Kirsch & Jungeblut, 1986, 1992). They were constructed as open-ended simulations, such that administration and scoring procedures were consistent with the procedures and criteria used to judge the successful performance of these tasks in workplace, daily-living, and educational settings (Kirsch et al., 1993).

Because the procedures used to validate the constructs and interpret the results of these adult literacy assessments proved so fruitful, we believed it would be useful to describe them here as one approach for validating a set of variables for the TOEFL 2000 test. Accordingly, we analyzed 20 reading items and 20 listening items from the experimental computer-based TOEFL item pool (Taylor et al., 1998). To identify the variables contributing to adults' reading and task difficulty in the prose and document domains, Kirsch and Mosenthal (Kirsch & Mosenthal, 1990a; Mosenthal & Kirsch, 1990d; Mosenthal, 1996, in press) began by modeling the processes required to complete prose and document tasks in the literacy assessments. This model is shown in Figure 7.

In the first step, readers identify a goal or purpose for searching and processing a text or document. In a test or an instructional situation, questions and directives determine the primary purpose for interacting with a text or document, and therefore also determine the information that readers must process in order to complete a cognitive activity. In open-ended tasks, the reader's goal is to identify information in the text that meets the conditions set forth in the question or directive. In multiple-choice tasks, the reader's goal is to identify information in the text that meets the conditions set forth in the question or directive and then to select the best choice from a list of options (Kirsch & Mosenthal, 1995).

It may be helpful to consider an example. Figure 8 presents a text from the computer-based TOEFL experimental reading set.

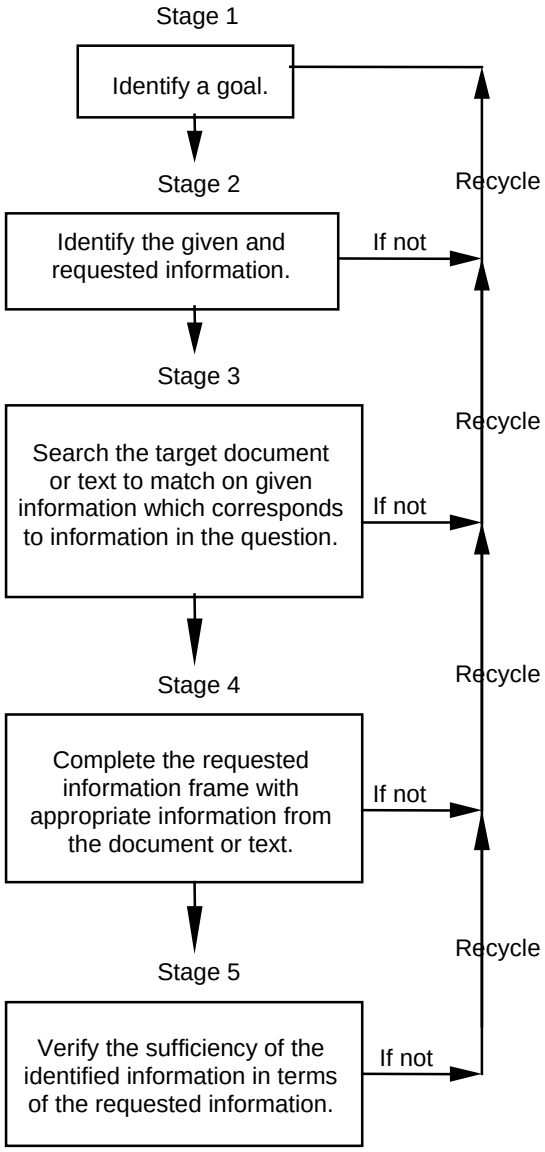

Figure 7. A model of prose and document processing in reading

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

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

Figure 8. "What Enabled Early Humans to Control the Use of Fire": sample text

Examinees were asked the following multiple-choice question as part of this set of reading tasks:

> Which of the following is the main topic of the passage?
>
> The positive effects of forest fires on early humans.
> Early indications of superior human intelligence.
> Characteristics that made it possible for early humans to control fire.
> Environmental conditions that threatened the survival of early humans.

Following the model shown in Figure 7, in the first step, the readers' purpose is to search the passage to identify the main idea.

In the second step, readers must distinguish between "given" and "requested" information in the question (Clark & Haviland, 1977; Mosenthal & Kirsch, 1991b; Kirsch & Mosenthal, 1992b). Given information is presumed to be true, and it conditions the requested information. Requested

information, on the other hand, is the specific information being sought. For example, in the preceding question based on the passage in Figure 8, the given information is that the text has a main idea. The requested information is information that summarizes the passage.

In the third step, readers must search and read (or read and search) a text or document to identify the necessary information that corresponds with information provided in the question and, in the case of multiple-choice items, in the list of choices. In carrying out this search, several matches may be tried before one or more adequate matches are achieved. If a literal or synonymous match is made between requested or given information and corresponding text or document information, readers may proceed to the next step. If such a match is not deemed adequate, readers may choose to make a match based on a low- or high-level text-based inference or on prior knowledge; or readers may recycle to the first step. Returning to the earlier example, readers may attempt to find a match between the choices and the text. In the process, they may make a literal match between the phrase "humans to control fire" in the text and the corresponding phrase in the third answer choice.

In the fourth step, readers complete the requested information frame by identifying the information asked for in the question. In some instances, readers are unable to complete the requested information frame based upon information associated with a current match for given information. In such cases, readers may recycle to an earlier step in the model, searching for information in another part of the text or document. In other instances, readers may once again need to make some sort of inference to relate the requested information to information in the text. In our previous example, once readers have matched on "humans to control (the use of) fire," they may select the overall choice—that is, "Characteristics that made it possible for early humans to control fire"—as the correct answer.

Finally, in the fifth step, readers may recycle to earlier steps to determine that all the conditions specified in a question have been adequately addressed. In some instances, readers may recycle in this step to identify information in different parts of a text or document or located elsewhere. In the earlier task, readers may read through the paragraph once again to ensure that the other choices do not represent the main point of the passage.

This test-taking model of reading can be applied to both documents and prose and to multiple-choice as well as open-ended tasks. Based on this model, we identified three domain continua as being among the best predictors of task difficulty. These continua (type of requested information, type of match, and plausibility of distractors), which were summarized in Section 4.4 of this paper, are elaborated in Appendices F, G, and H.

The next question of interest is, how useful are these variables in accounting for examinee performance on the experimental TOEFL CBT reading tasks (Taylor et al., 1998)? To answer this question, we analyzed 20 CBT reading tasks using the procedures described in Appendices F through H. The results are shown in Table 1. Using "percentage correct scores," we regressed item difficulty on the value of the variables. The variables predicted 86 percent of the variance in task difficulty on the reading tasks.

Table 1
Analysis of TOEFL CBT Reading Tasks Using Variable Descriptors and Values

Task No.	Content	p-value	Type of Information	Type of Match	Plausibility of Distractors
1	Fire	.67	theme = 5	locate = 1	3
2	Fire	.49	ambiguity = 5	locate + 2 phrases + 1 low-level inference = 3	5
3	Fire	.68	pronoun referent = 3	locate + antecedent = 3	3
4	Fire	.27	equivalence = 5	locate + syntagmatic relation + prior knowledge = 7	5
5	Fire	.55	pronoun referent = 3	locate + antecedent = 3	5
6	Fire	.30	difference = 5	integrate + contrast + antecedent = 6	5
7	Fish	.70	verification = 3	cycle + between paragraphs = 3	2
8	Fish	.88	goal = 3	locate = 1	3
9	Fish	.70	goal = 3	locate = 1	5
10	Fish	.85	equivalence = 5	locate = 1	2
11	Fish	.58	problem = 3	integrate + identify condition = 6	4
12	Fish	.61	verification = 3	integrate + between paragraphs = 4	3
13	Fish	.85	action = 2	integrate = 3	2
14	Quartz	.83	attribute = 2	integrate = 3	2
15	Quartz	.70	explanation = 4	locate + 2 phrases + low-level inference for new information = 4	2
16	Quartz	.56	attribute = 2	integrate + syntagmatic = 6	5
17	Quartz	.52	explanation = 4	locate + high-level inference for new information = 5	5
18	Quartz	.72	equivalence = 5	locate = 1	2
19	Quartz	.55	equivalence = 5	locate + high-level inference for new information = 5	5
20	Quartz	.17	theme = 5	integrate + between paragraphs + high-level inference for new information = 8	3

Because listening and reading tasks have some shared characteristics, we thought that it would also be useful to apply the model to 20 listening items from the experimental computer-based TOEFL test. Table 2 presents the analyses for these tasks. The fit for the listening tasks was not quite as good as it was for the reading tasks, accounting for 79 percent of the variance in task difficulty. Further, type of information played an insignificant role in the multiple regression in listening. This suggests that, while there is some overlap in the reading and listening

constructs, at least with the limited sample of items analyzed, there are variables that may be unique to each modality.

These findings suggest that a set of variables can be identified that are both specific to and generalizable across the four modalities, and that methodologies can be applied to account for significant percentages of task difficulty. In the next section, we continue to borrow from adult literacy research to show how this information can be used to interpret and describe what is being measured along a particular scale. We also include an example from Australia where a scheme was included as part of the development of a test of Chinese.

Table 2
Analysis of TOEFL CBT Listening Tasks Using Variable Descriptors and Values

Task No.	Content	p-value	Type of Information	Type of Match	Plausibility of Distractors
1	Dialog	.74	assertion = 3	locate + inference = 3	2
2	Dialog	.68	verification = 3	locate + inference = 3	3
3	Short conver-	.96	goal = 3	locate = 1	2
4	sational set	.82	action = 2	locate = 1	3
5	Academic	.76	goal = 3	locate + inference = 2	3
6	discussion/	.40	similarity = 4	integrate = 3	5
7	linguistics	.30	manner = 3	integrate + number of phrases/multiple response = 7	5
8		.47	conditions = 3	cycle + inference/multiple response = 6	5
9	Mini talk/	.92	thing = 1	locate = 1	5
10	geology	.68	problem = 4	cycle + multiple response = 3	2
11		.61	location = 2	integrate = 3	4
12	Mini talk/	.73	location = 2	integrate = 3	4
13	art	.72	goal = 3	locate + inference = 3	2
14		.55	cause = 4	cycle + multiple response = 3	2
15		.76	goal = 3	locate = 1	5
16		.67	similarity = 4	integrate + multiple response = 4	2
17	Mini talk/	.75	goal = 3	locate + number of phrases = 3	2
18	botany	.53	kind/type = 2	cycle + number of phrases/ multiple response = 4	5
19		.52	sequence = 3	cycle + number of phrases/ multiple response = 7	1
20		.65	type = 2	locate + inference = 3	3

4.6 Building an Interpretive Scheme

Identifying and validating a set of variables provides a basis not only for guiding item writing and test construction but also for developing an interpretive scheme. This section discusses some of the ways in which an interpretive scheme can be built.

The procedures proposed here are not new. They derive from Beaton's anchored proficiency procedures (Beaton & Allen, 1992; Messick, Beaton, & Lord, 1983), but are more flexible than originally developed by Beaton and as used in the early National Assessment of Educational Progress analyses. They have been used in numerous adult literacy studies in the United States (Kirsch, 1995), and Brown, Elder, Lumley, McNamara, and McQueen (1992) have used them in Australia to establish levels in a test of Chinese. The Brown et al. application is included here because it addresses a second language setting and thus is directly relevant to the TOEFL test.

The procedures described here rely on item response theory (IRT) analyses, which make it possible to place tasks and examinees on the same scale—tasks at the location corresponding to their difficulty, and examinees at the location corresponding to their abilities. Figure 9 provides an example. Tasks with scale values above an examinee's score (see the position of task 1 relative to examinee A's score) are those that the individual has a low probability of performing successfully. Conversely, tasks with scale values below the examinee's score (see the position of task 2 relative to examinee A's score) are those that the individual has a high probability of answering correctly. Thus, IRT makes it possible to characterize individuals' abilities in terms of the tasks above and below their ability levels. For example, in the TOEFL 2000 context, if task 1 were set in an academic classroom milieu and asked the examinee to read a loosely structured text, then examinee A would be described as someone who had a low probability of performing this task successfully.

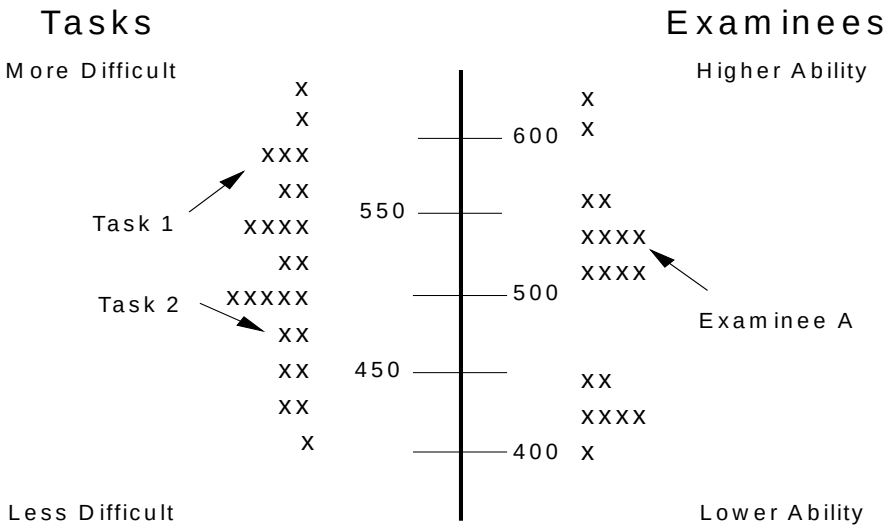

Figure 9. The relation between task difficulty and examinee ability on the same IRT scale

Note: The scale was transformed to the traditional TOEFL score range. Data were constructed for illustrative purposes.

More precisely, because the task scores can be interpreted as probabilities of success relative to a particular ability, it is possible to characterize an examinee's score as the probability of answering tasks at any point on the ability scale. Thus, for example, if task 1 represented a criterion task, examinee A could be described by his or her probability of responding correctly to that task. Of course, we are seldom interested in the specific items on the test. Each item is intended to be representative of a class of potential items with similar characteristics, and these in turn are intended to be representative of tasks that individuals have to perform outside of the testing situation. In short, we need to generalize from the particular items on a TOEFL test to the set of academic tasks that are the real area of concern.

In some cases, tasks with similar scale scores tend to have similar properties, and these properties can be recorded using sets of variables. This was the case for the U.S. and international literacy assessments and for the Chinese test in Australia. Rather than just locate individual test tasks on the scale, then, one can define task characteristics at various points on the scale. This makes it possible to describe an examinee's ability in terms of the characteristics of tasks whose difficulty approximates his or her estimated ability. Further, one can establish levels of performance along the scale. Figure 10 shows the clusters found by Brown et al. in their analysis of the Chinese test. Although these researchers chose to describe the levels in terms of examinees' abilities, they could just as easily have phrased them in terms of task characteristics (e.g., "A task at this level requires examinees to . . .").

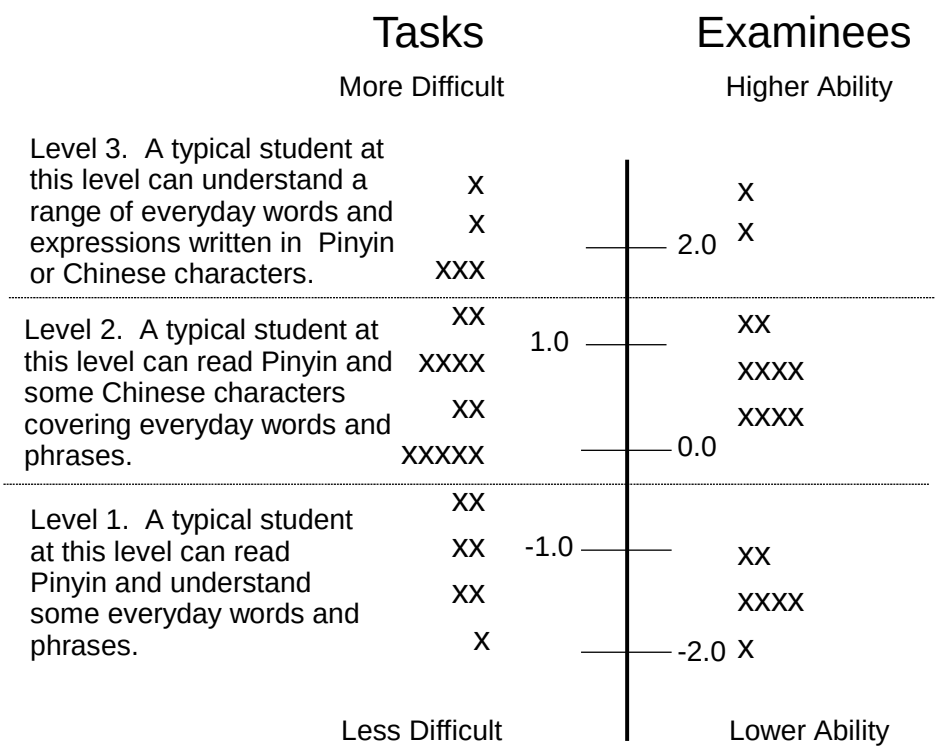

Figure 10. Levels for the Chinese test mapped onto the IRT scale display

Adapted from Brown et al. (1992).

In the case of the TOEFL 2000 test, we hope to identify sets of variables, validate some of these variables, and use these to establish an interpretive scheme. This would allow test developers and test users to generalize beyond the test items to other tasks that are of interest and that can be characterized in the terms of the framework. In this way, the potential ability of TOEFL examinees to deal with common university tasks could be estimated by comparing the examinee's score with the placement of such tasks on the TOEFL scale. The scale score range for each level is not set prior to the analysis, but flows from it. Once set in the norming study, however, the score ranges that define the levels can be maintained in new operational forms through standard equating procedures.

Such procedures were followed for the literacy assessments identified earlier (Kirsch et al., in press). In analyzing the results of these assessments, tasks were placed on the literacy scales using an 80 percent response probability criterion. In other words, a task's scale value was the point on the scale at which respondents had an 80 percent probability of performing the task successfully. To expand the interpretability of the survey data, Kirsch et al. (1991, 1993) then

divided the literacy scales into five levels. These levels, shown in Table 3, had the following ranges: Level 1 included tasks with scores at or below 225; Level 2 included tasks with scores ranging from 226 to 275; Level 3 included tasks ranging from 276 and 325; Level 4 included tasks with scores ranging from 326 to 375; and Level 5 included tasks with scores above 375.

To determine the internal consistency of the variable scores within these levels, Kirsch et al. (1991, 1993) qualitatively determined score ranges. This was accomplished by examining the relative values for the process variables (type of information, type of match, and plausibility of distractors) within each level and then identifying ranges that were unique to these levels. The resulting score ranges are shown in Table 3. For instance, it was noted that Level 1 ideally would include some variable values of 1, 2, and 2 or lower. At Level 4, the three process variables would have values of 4, 4 or higher, and 4 or lower. At Level 5, the three process variables were expected to assume values of 5, 5 or higher (for type of match), and 5 or lower.

Table 3
Internal Consistency of Variable Scores by Level for Prose Tasks in Five Adult Literacy Surveys

	Level 1 (RP80: ≤225)	Level 2 (RP80: 226-275)	Level 3 (RP80: 276-325)	Level 4 (RP80: 326-375)	Level 5 (RP80: >375)	
Variable score ranges	1, 1, 2 or lower	2, 2, 2; or 3, 3 or lower, 3 or lower	4, 3 or lower, 3 or lower	4, 4 or higher, 4 or lower	5, 5 or higher, 5 or lower	
Percentage and raw number of tasks within level with combination scores consistent with variable score range	88% (7 out of 8 tasks)	81% (26 out of 32 tasks)	85% (60 out of 71 tasks)	90% (36 out of 40 tasks)	86% (12 out of 14 tasks)	Overall 85% (141 out of 165 tasks)

As a next step, the variable scores for the tasks in each level were compared with the criteria scores for that level. Tasks with variable scores that met the range score criteria were said to be "internally consistent" within the level. Thus, if a task had a type of information score of 1, a type of match score of 2, and a plausibility of distractors score of 1, it met the range-score criteria for Level 1 and was said to be internally consistent.

For the five adult literacy surveys, overall internal consistency within levels was determined by dividing the number of prose tasks that met a level's range score criteria by the total number of tasks in the level (Mosenthal, in press). As shown in Table 3, on the prose scale, the percentages of internally consistent tasks in each level ranged from 81 percent (Level 2) to 90 percent (Level 4). Of the 165 unique tasks in the five surveys, 85 percent (or 141) were internally consistent. Of the 24 tasks that were not, 19 failed to meet the range score criteria because a single variable value was one point higher or lower than the criteria. The remaining five tasks also differed in terms of a single variable value, but for these tasks the value was two points above or below the criteria.

Overall, then, the variable score ranges within the levels were highly consistent. This consistency makes it possible to specify constructs that tend to be highly characteristic of task difficulty and reader proficiency in each of the five levels, as shown in Table 4.

Table 4
Constructs Underlying Prose Task Difficulty and Examinee Proficiency by Level in Five Adult Literacy Surveys

Level 1 (≤ 225) Most of the tasks in this level require readers to identify information which is quite concrete, including a person, place, or thing, or an attribute, amount, type, temporal, action, procedure, or location. To complete these tasks, readers must process relatively short text to locate a single piece of information which is identical to (or synonymous with) the information given in the question or directive. If distractors appear in the text, they tend to be located in a paragraph other than the one in which the correct answer occurs.
Level 2 (226 to 275) Like tasks in Level 1, many tasks in Level 2 ask readers to identify information which is fairly concrete. In this level, however, some tasks also require readers to identify information representing manner, goal, purpose, attempt, alternative, or condition. Level 2 tasks often require readers to make a low-level text-based inference or identify a condition or antecedent in order to identify requested information in a text. Most tasks in this level have a distractor for given or requested information, but these are not typically found in the same paragraph as the answer.
Level 3 (276 to 325) Tasks in Level 3 tend to require readers to identify conditional information, or to indicate a reason or explanation. Level 3 tasks often require readers to make literal, synonymous, or low-level inference matches between the question or directive and the text. Unlike the Level 1 and 2 tasks, Level 3 tasks usually require readers to identify and list multiple responses, the number of which is specified in the question or directive. The questions and directives for the tasks in this level also tend to consist of several phrases. The tasks generally require readers to complete requested information by identifying special conditional information stated in the question or directive or by establishing antecedence between a pronoun and its referent. Distractors for both given and requested information tend to be present; typically, these distractors appear in different paragraphs from one another, and neither appears in the same paragraph as the answer.
Level 4 (326 to 375) Tasks in this level tend to require readers to identify rather abstract information, including reason, evidence, explanation, causation, result, comparison, and contrast. In terms of type of match, Level 4 tasks generally require readers not only to locate information, but also to cycle and integrate. Again, multiple responses may be required, but the number of responses is not specified. Like Level 3 tasks, Level 4 tasks often require readers to complete requested information by identifying special conditional information stated in the question or directive, or by establishing antecedence between a pronoun and its referent. In other cases, examinees must make high-level text-based inferences to distinguish the correct requested information from distracting information. Distractors for both given and requested information tend to be present; both types of distractors may appear in the same paragraph as the answer.
Level 5 (≥ 376) Tasks in this level tend to require readers to identify quite abstract information, including contrast, equivalence, and theme (or summary). In terms of type of match, Level 5 tasks often require readers not only to locate, cycle, and integrate, but also to generate information. Specialized prior knowledge may be required to complete the requested information. Distractors for both given and requested information are almost always present in Level 5 tasks; both types of distractors generally appear in the same paragraph as the answer.

Although we have identified and constructed levels that are consistent across the various adult literacy surveys, the range values were based on open-ended responses. Accordingly, to the extent that the TOEFL 2000 test will use a combination of open-ended and multiple-choice tasks to assess listening and speaking abilities, our score ranges, as well as the constructs characterizing different levels of proficiency, task difficulty, and production complexity, will have to be empirically determined anew, especially because the examinees are adults who are nonnative speakers of English. The constellation of constructs by levels will also have to be reconfigured. Since academic tasks are most likely to fall within Levels 3, 4, and 5, there is no doubt that the construct specification for these levels needs to be significantly enhanced to more precisely define the depth and breadth of the reading, listening, writing, and speaking tasks at these levels. In keeping with the reading and listening tasks, it is our intent to characterize writers' and speakers' productions in terms of levels of proficiency, task difficulty, and response complexity.

It is important to keep in mind that some of the variables identified in the adult literacy research may not turn out to be relevant in defining the scale levels for the TOEFL 2000 test. Once the relevant variables are identified through an empirical process, they can be used to generalize from the TOEFL 2000 measures to test users' criteria. This does not mean that the other variables are necessarily irrelevant; they too must be considered. However, they serve to bound the domain rather than to locate points within the domain. Score users could not safely generalize TOEFL 2000 test results to work situations, for example, if all of the tasks were drawn from academic settings.

The empirical studies could show that the existing set of variables is inadequate. In other words, it is possible that the data on examinee performance will not yield any clusters of tasks that can be described by these variables, or that the set will not be adequate to describe differences in examinee performance. Thus, the framework itself is subject to empirical evaluation, and the pre-operational trials of the TOEFL 2000 instruments derived from the framework must be seen as a vital part of this evaluation process.

In closing this section, it is important to outline some of the advantages associated with building an interpretive scheme—that is, of identifying and validating the construct levels within a task domain. One advantage is that testing and instruction do not have to be "competency based," whereby each task in a domain is tested and taught as a unique competency (Mosenthal & Kirsch, 1989a). Instead, levels can be used to define a finite number of proficiencies that underlie all possible tasks associated with a domain. These tasks may currently exist in the domain or may be incorporated at a future time.

Specifying test and instructional tasks in terms of construct-level characteristics instead of task characteristics also reduces the number of instructional tasks needed to enhance students' proficiency in the domain. Moreover, this approach optimizes transfer in that it focuses testing and instruction on underlying task construct characteristics rather than on superficial features (Embretson, 1983, 1993; Nichols, 1994; Nitko, 1989).

Knowing what tasks are likely to fall within a domain based on variable and construct specifications also enables test developers to write tasks that address precise testing and instructional purposes. This eliminates the hit-or-miss strategy of having to continually calibrate the difficulty of tasks (and associated item difficulty parameters) through repeated field testing.

Finally, to the extent that tasks within a level share similar empirically determined construct characteristics, the levels can be used to define students' zones of proximal proficiency. Ideally, these zones reflect a relatively consistent range of variable values that, in turn, reflect a rather restricted range of task construct characteristics. Taken together, this range of variables and construct characteristics define an empirically determined and precisely specified set of "subdomains" whose constructs characterize student proficiency, task difficulty, and response complexity in terms of a proficiency/difficulty/complexity hierarchy.

5. Plans for Proceeding

From our perspective, it is important to reach consensus about what the TOEFL 2000 test will measure and how it will be delivered to the field in the next century. To accomplish this goal, we plan to build consensus through formal and informal means, beginning with internal workshops and reviews and then expanding to broader, external circles of individual consultants, TOEFL committees, and larger professional meetings. In December 1996, we presented an initial draft of the framework to ETS test developers, researchers, and program direction staff, many of whom will share responsibility for constructing and delivering a reliable and valid TOEFL 2000 instrument. A few days later, we presented the same framework to a small group of consultants who had served on TOEFL committees and produced monographs on various aspects of the TOEFL 2000 project. Both groups, internal staff and external consultants, were then invited to provide written reviews of the framework. The current framework incorporates the feedback from these first two rounds of review.

Continuing the cycle of review and feedback, the next step was to share the framework with members of the TOEFL Committee of Examiners and Research Committee. Their comments and ideas were considered when the framework was subsequently taken to the TOEFL Policy Council for review in May 1997. The resulting iteration of the framework was then presented at professional meetings and test user focus groups to gauge public reactions and continue building consensus. In addition to making this framework paper available to the public upon request, we propose submitting TOEFL 2000 research reports and papers for journal publication as a means of ensuring wider dissemination and encouraging extensive discussion of the framework and research on the project.

As this document is disseminated more widely, individuals both within and outside of ETS will need to join the current team in assuming responsibility for the continued development of this work. Four working teams were created and charged with (a) using the current framework to operationalize specific frameworks for reading, writing, listening, and speaking, and (b) developing a research agenda to support the framework. The activities of these framework teams are being organized and directed by ETS research and test development staff. Research and test development collaboration is crucial to ensuring the creation of frameworks that both have an empirical basis and provide adequate specifications from which forms of the new test can be generated. ETS staff, as well as outside experts, will participate in carrying out the research agenda and conducting an on-going program of research as the new test is implemented. In addition to playing a key role in defining and operationalizing the new test, test development staff will eventually be responsible for training other test developers and item writers and moving the test model to a production mode. External members of these teams will include experts from various disciplines, such as first and second language teaching, instructional design, technology applications, and language testing research.

It is expected that the research agenda that emerges with the frameworks will identify and prioritize the key issues that must be addressed over the next several years as the TOEFL 2000 program moves from a research and development project to an operational test.

The project will proceed with the continued oversight of the TOEFL Policy Council and advice of the TOEFL Committee of Examiners, TOEFL Research Committee, and other specialists with whom project teams will consult. It is important to reiterate that this framework is a work in progress and is expected to be informed and refined by research and on-going dialog with TOEFL constituencies.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*: 191-204.

Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives. Handbook I: Cognitive domain*. New York: David McKay.

Brown, A., Elder, C., Lumley, T., McNamara, T., & McQueen, J. (1992, February). Mapping abilities and skill levels using Rasch techniques. Paper presented at Language Testing Research Colloquium, Vancouver, BC.

Brown, H. D. (1987). Principles of language learning and teaching (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Brown, P., & Levinson, S. (1978). Universals in language usage: Politeness phenomena. In E. N. Goody (Ed.), *Questions and politeness*. Cambridge: Cambridge University Press.

Canale, M. (1983). On some dimensions of language proficiency. In J. Oller (Ed.), *Issues in language testing research* (pp. 333-342). Rowley, MA: Newbury House.

Canale, M. (1988). The measurement of communicative competence. *Annual Review of Applied Linguistics, 8,* 67-84.

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

Carroll, J. B. (1961). Fundamental considerations in testing for English proficiency of foreign students. *Testing the English proficiency of foreign students.* Washington, DC: Center for Applied Linguistics, 31-40.

Cazden, C. B. (1996, March). *Communicative competence: 1966-1996.* Paper presented at the meeting of the American Association of Applied Linguistics, Chicago, IL.

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

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.), *Discourse production and comprehension* (pp. 1-39). Norwood, NJ: Ablex.

Coleman, A. (1934). A survey of tendencies in modern language teaching, 1927-33: Retrospect and prospect. In A. Coleman (Ed.), *Experiments and studies in modern language teaching* (pp. 50-99). Chicago: University of Chicago Press.

Crystal, D. (1991). *A dictionary of linguistics and phonetics*. Cambridge, MA: Basil Blackwell, Inc.

Davies, A. (1987). Certificate of Proficiency in English. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 20-21). Washington, DC: Teachers of English to Speakers of Other Languages.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations.* (TOEFL Monograph Series Report No. 8). Princeton, NJ: Educational Testing Service.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. E. (1985). *TOEFL from a communicative viewpoint on language proficiency: A working paper.* (TOEFL Research Report No. 17). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1991). *ETS tests of applied literacy skills: Prose, document, and quantitative.* New York: Simon and Schuster.

Educational Testing Service. (1994). *Computer-based testing benefits*. Princeton, NJ: Author.

Educational Testing Service. (1996). Shaping TOEFL's future: Computer-based testing. *TOEFL Update: 1996 Edition*. [Brochure]. Princeton, NJ: Author.

Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). *Development of a scale for assessing the level of computer familiarity of TOEFL examinees.* (TOEFL Research Report No. 60). Princeton, NJ: Educational Testing Service.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ: Erlbaum.

Ervin-Tripp, S. M. (1972). Sociolinguistic rules of address. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 225-40). Harmondsworth: Penguin.

Francis, W. N., Kucera, H., & Mackie, A. W. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Freedle, R., & Kostin, I. (1993). *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items.* (TOEFL Research Report No. 44). Princeton, NJ: Educational Testing Service.

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

Gough, D. (1995). *Trends in international admission.* TOEFL 2000 Internal Report. Princeton, NJ: Educational Testing Service.

Grosse, C. U. (1991). The TESOL methods course. *TESOL Quarterly, 25,* 29-50.

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

Halliday, M. A. K. (1970). Language structure and language function. In J. Lyons (Ed.), *New Horizons in Linguistics* (pp. 140-165). Harmondsworth: Penguin.

Halliday, M. A. K. (1973). *Explorations in the functions of language*. London: Arnold.

Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. London: Longman.

Hamp-Lyons, L. (1987). Cambridge First Certificate in English. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 18-19). Washington, DC: Teachers of English to Speakers of Other Languages.

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

Handschin, C. (1923). *Methods of teaching modern languages.* Yonkers-on-Hudson, NY: World Book Company.

Heath, S. B. (1980). The functions and uses of literacy. *Journal of Communications, 30,* 123-133.

Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000.* (TOEFL Monograph Series Report No. 4). Princeton, NJ: Educational Testing Service.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics.* Honolulu, HI: Second Language Teaching and Curriculum Center, University of Hawaii.

Hymes, D. H. (1971). Competence and performance in linguistic theory. In R. Huxley & E. Ingram (Eds.), *Language acquisition: Models and methods*. London: Academic Press.

Hymes, D. H. (1972a). Models of the interaction of language and social life. In J. J. Gumperz & D. H. Hymes (Eds.), *Directions in sociolinguistics* (pp. 35-71). New York: Holt, Rinehart, & Winston.

Hymes, D. H. (1972b). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-93). Harmondsworth: Penguin.

Hymes, D. H. (1974). *Foundations in sociolinguistics.* Philadelphia: University of Pennsylvania Press.

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

Kirsch, I. (1995). Literacy performance on three scales: Definitions and results. In *Literacy, Economy and Society: Results of the first international adult literacy survey* (pp. 27-53). Paris: Organisation for Economic Co-operation and Development (OECD).

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees*. (TOEFL Research Report No. 59). Princeton, NJ: Educational Testing Service.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults* (NAEP Report No. 16-PL-01). Princeton, NJ: Educational Testing Service.

Kirsch, I. S., & Jungeblut, A. (1992). *Profiling the literacy proficiencies of JTPA and ES/UI populations: Final report to the Department of Labor*. Princeton, NJ: Educational Testing Service.

Kirsch, I. S., Jungeblut, A., & Campbell, A. (1991). *The ETS tests of applied literacy skills* (*Document literacy, prose literacy, and quantitative literacy*): *Administration and scoring manual.* Princeton, NJ: Educational Testing Service.

Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey*. Washington, DC : U.S. Department of Education.

Kirsch, I., Jungeblut, A., & Mosenthal, P. B. (in press). Moving toward the measurement of adult literacy. In *National Adult Literacy Survey Technical Report*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kirsch, I. S., & Mosenthal, P. B. (1989). Building documents by combining simple lists. *Journal of Reading, 33*, 132-135.

Kirsch, I. S., & Mosenthal, P. B. (1990a). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 25*, 5-30.

Kirsch, I. S., & Mosenthal, P. B. (1990b). Mimetic documents: Pictures. *Journal of Reading, 34*, 216-220.

Kirsch, I. S., & Mosenthal, P. B. (1990c). Nested lists. *Journal of Reading, 33*, 294-297.

Kirsch, I. S., & Mosenthal, P. B. (1990d). Understanding forms (Part I). *Journal of Reading, 33*, 542-545.

Kirsch, I. S., & Mosenthal, P. B. (1990e). Understanding forms (Part II). *Journal of Reading, 33*, 636-641.

Kirsch, I. S., & Mosenthal, P. B. (1990/1991). Mimetic documents: Diagrams. *Journal of Reading, 34*, 290-294.

Kirsch, I. S., & Mosenthal, P. B. (1991). Understanding definitions, descriptions, and comparison/contrasts. *Journal of Reading, 35*, 156-160.

Kirsch, I. S., & Mosenthal, P. B. (1991/1992). Understanding cases and classes through knowledge modeling. *Journal of Reading, 35*, 332-338.

Kirsch, I. S., & Mosenthal, P. B. (1992a). Understanding process knowledge models. *Journal of Reading, 35*, 490-497.

Kirsch, I. S., & Mosenthal, P. B. (1992b). How to navigate a document using locate, known/need-to-know strategies. *Journal of Reading, 36*, 140-144.

Kirsch, I. S., & Mosenthal, P. B. (1992/1993). Integration strategies: Higher order thinking applied to documents. *Journal of Reading, 36*, 322-327.

Kirsch, I. S., & Mosenthal, P. B. (1995). Interpreting the IEA reading literacy scales. In M. Binkley, K. Rust, & M. Winglee (Eds.), *Methodological issues in comparative educational studies: The case of the IEA reading literacy study* (pp. 135-192). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Lannon, J. M. (1982). *Technical writing* (2nd ed.). Boston: Little, Brown.

Long, M., & Richards, J. (Eds.). (1987). *Methodology in TESOL, a book of readings.* New York: Newbury House.

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

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*, 9-20.

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

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Report No. 83-1). Princeton, NJ: Educational Testing Service.

Mills, G. H., & Walter, J. A. (1978). *Technical writing* (4th ed.). New York: Holt, Rinehart and Winston.

Morrow, K. (1987). University of Cambridge Local Examinations Syndicate: Preliminary English Test. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 21-22). Washington, DC: Teachers of English to Speakers of Other Languages.

Mosenthal, P. B. (1976-1977). Bridge principles in an abridged reply to Goodman. *Reading Research Quarterly*, *12*, 586-603.

Mosenthal, P. B. (1982). Toward a paradigm of children's writing classroom competence. In B. A. Hutson (Ed.), *Advances in reading/language research: A research annual* (Volume 1, pp. 125-154).

Mosenthal, P. B. (1983). Defining classroom writing competence: A paradigmatic perspective. *Review of Educational Research, 53*, 2, 217-251.

Mosenthal, P. B. (1985). Defining the expository discourse continuum. *Poetics, 15*: 387-414.

Mosenthal, P. B. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, 88,* 314-332.

Mosenthal, P. B. (in press). Defining prose task characteristics for use in computer-adaptive testing and instruction. *American Educational Research Journal*.

Mosenthal, P. B., & Kirsch, I. S. (1989a). Designing effective adult literacy programs. *Poetics, 18,* 239-256.

Mosenthal, P. B., & Kirsch, I. S. (1989b). Intersecting lists. *Journal of Reading, 33,* 210-213.

Mosenthal, P. B., & Kirsch, I. S. (1989c). Lists: The building blocks of documents. *Journal of Reading, 33,* 58-60.

Mosenthal, P. B., & Kirsch, I. S. (1990a). Understanding general reference maps. *Journal of Reading, 34,* 60-63.

Mosenthal, P. B., & Kirsch, I. S. (1990b). Understanding graphs and charts (Part I). *Journal of Reading, 33,* 371-373.

Mosenthal, P. B., & Kirsch, I. S. (1990c). Understanding graphs and charts (Part II). *Journal of Reading, 33,* 454-457.

Mosenthal, P. B., & Kirsch, I. S. (1990d). Understanding thematic maps. *Journal of Reading, 34,* 136-140.

Mosenthal, P. B., & Kirsch, I. S. (1991a). Mimetic documents: Process schematics. *Journal of Reading, 34,* 390-397.

Mosenthal, P. B., & Kirsch, I. S. (1991b). More mimetic documents: Procedural schematics. *Journal of Reading, 34,* 486-490.

Mosenthal, P. B., & Kirsch, I. S. (1991c). Information types in nonmimetic documents: A review of Biddle's wipe-clean slate. *Journal of Reading, 34,* 654-660.

Mosenthal, P. B., & Kirsch, I. S. (1991d). Toward an explanatory model of document process. *Discourse Processes, 14,*147-180.

Mosenthal, P. B., & Kirsch, I. S. (1991e). Extending prose comprehension through knowledge modeling. *Journal of Reading, 35,* 58-61.

Mosenthal, P. B., & Kirsch, I. S. (1991f). Using knowledge models to understand steady states. *Journal of Reading, 35,* 250-255.

Mosenthal, P. B., & Kirsch, I. S. (1992a). Cycle strategies in document search: From here to there to wherever. *Journal of Reading, 36*, 238-242.

Mosenthal, P. B., & Kirsch, I. S. (1992b). Understanding the knowledge models of simple events. *Journal of Reading, 35*, 408-415.

Mosenthal, P. B., & Kirsch, I. S. (1992c). Understanding knowledge acquisition from a knowledge model perspective. *Journal of Reading*, *35*, 588-596.

Mosenthal, P. B., & Kirsch, I. S. (1993a). Generate strategies: Coping without cues and clues. *Journal of Reading, 36*, 416-419.

Mosenthal, P. B., & Kirsch, I. S. (1993b). Profiling students' document strategy abilities. *Journal of Reading, 36*, 578-583.

Munby, J. (1978). *Communicative syllabus design.* Cambridge: Cambridge University Press.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575-603.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 447-474). New York: Macmillan.

Oller, J. W. (1979). *Language tests at school.* London: Longman.

Pearson, P. D., & Johnson, D. (1978). *Teaching reading comprehension.* New York: Holt, Rinehart, and Winston.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction and research. In G. Crookes & S. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 9-34). Philadelphia: Multilingual Matters Ltd.

Pickett, N. A., & Laster, A. A. (1975). *Technical English* (2nd ed.). San Francisco: Canfield Press.

Powell, W. (in press). *Looking back, looking forward: Trends in intensive English program enrollments.* (TOEFL Monograph Series Report No. 14). Princeton, NJ: Educational Testing Service.

Rea, P. M. (1987). Test of English for Educational Purposes. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 77-79). Washington, DC: Teachers of English to Speakers of Other Languages.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversations. *Language, 50*: 696-735.

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

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica,* 8: 289-327.

Selinker, L., Tarone, E., & Hanzeli, V. (Eds.). (1981). *English for technical and academic purposes: Studies in honor of Louis Trimble*. Rowley, MA: Newbury House.

Skehan, P. (1995). A framework for the implementation of task-based instruction. *Applied Linguistics, 17,* 1: 38-62.

Stansfield, C. W. (Ed.). (1986). *Toward communicative competence testing: Proceedings of the second TOEFL invitational conference.* (TOEFL Research Report No. 21). Princeton, NJ: Educational Testing Service.

Tarone, E. (1983). On the variability of interlanguage systems. *Applied Linguistics, 4*: 142-63.

Taylor, C. (1993). *Report of TOEFL score users focus groups.* TOEFL 2000 Internal Report. Princeton, NJ: Educational Testing Service.

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

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks.* (TOEFL Research Report No. 61). Princeton, NJ: Educational Testing Service.

The New London Group: Cazden, C., Cope, B., Fairclough, N., Gee, J., Luke, A., Michaels, S., & Nakata, M. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review, 1, 60-92.*

Tony, T. (1987). Association of Recognised Language Schools' Oral Examinations in Spoken English. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 5-7). Washington, DC: Teachers of English to Speakers of Other Languages.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed-response, performance testing, and portfolio assessment* (pp. 29-44). Hillsdale, NJ: Erlbaum.

van Ek, J. A., & Alexander, L. G. (1975). *Threshold level English.* Oxford: Pergamon.

van Els, T. J. M., & Engels, L. K. (Eds.). (1983). *Notional-functional syllabus in language learning.* Uitgerverij, Amsterdam: VU Bookhandel.

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

Weir, C. J. (1990). *Communicative language testing.* NY: Prentice Hall.

Wesche, M. B. (1987). Communicative testing in a second language. In M. H. Long & J. C. Richards (Eds.), *Methodology in TESOL: A book of readings* (pp. 373-394). New York: Newbury House.

Appendix A. TOEFL Monograph Series

Monographs related to trends in international student enrollments:

Powell, W. (in press). *Looking back, looking forward: Trends in intensive English program enrollments.* (TOEFL Monograph Series Report No. 14). Princeton, NJ: Educational Testing Service.

Monographs related to test constructs:

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

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

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations.* (TOEFL Monograph Series Report No. 8). Princeton, NJ: Educational Testing Service.

Douglas, D. & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revisions project.* (TOEFL Monograph Series Report No. 9). Princeton, NJ: Educational Testing Service.

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

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

Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000.* (TOEFL Monograph Series Report No. 4). Princeton, NJ: Educational Testing Service.

Lazaraton, A. & Wagner, S. (1996). *The revised TSE: Discourse analysis of native and nonnative speaker data*. (TOEFL Monograph Series Report No. 7). Princeton, NJ: Educational Testing Service.

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

Monographs related to psychometric models and procedures:

Carey, P. A. (1996). *A review of psychometric and consequential issues related to performance assessment*. (TOEFL Monograph Series Report No. 3). Princeton, NJ: Educational Testing Service.

Tang, L. (1996). *Polytomous IRT models and their applications in large-scale testing programs: Review of literature*. (TOEFL Monograph Series Report No. 2). Princeton, NJ: Educational Testing Service.

Monographs related to technology applications:

Burstein, J. C., Kaplan, R. M., Rohen-Wolff, S., Zuckerman, D. I., & Lu, C. (1999). *A review of computer-based speech technology for TOEFL 2000.* (TOEFL Monograph Series Report No. 13). Princeton, NJ: Educational Testing Service.

Frase, L., Gong, B., Hansen, E., Kaplan, R., Katz, I., & Singley, K. (1998). *Technologies for language testing*. (TOEFL Monograph Series Report No. 11). Princeton, NJ: Educational Testing Service.

Hansen, E., & Willut, C. (1998). *Computer and communications technologies in colleges and universities in the year 2000.* (TOEFL Monograph Series Report No. 12). Princeton, NJ: Educational Testing Service.

Appendix B. Description of the "Purpose" Variable

The following descriptions are adapted from Halliday (1973, pp. 3-38) and Brown (1987, p. 203). Each description is followed by a sample task.

The *heuristic* purpose involves language used to acquire knowledge, to learn about things. A student's reading a chapter in an assigned text is an example. Another example is a student's listening to a lecture.

The *instrumental* purpose serves to manipulate the environment, to cause certain events to happen; language is used as a means for getting things done. Copying down the call number of a book one wants to check out of the library is an example of an instrumental function.

The *regulatory* purpose of language refers to directing the behavior of others. The regulation of encounters among people—approval, disapproval, behavior control, setting laws and rules—are all regulatory features of language. The following text taken from a university's course registration form is regulatory: "Upon completion of this form: 1. Obtain advisor's signature; 2. Take to department of major for stamp."

The *personal* purpose allows a speaker to express feelings, emotions, personality, "gut-level" reactions. A student expressing worry about an upcoming test is one example of the personal function. Another example is two students talking about the progress of a fellow student.

The *representational* purpose refers to expressing propositions; it is the use of language to make statements, convey facts and knowledge, explain, or report—that is, to "represent" reality as one sees it. Telling someone the time is an example.

The *interactional* purpose of language serves to ensure social maintenance; it is the communicative contact between and among human beings that simply allows them to keep channels of communication open. An example is thanking a reference librarian for his or her help. Talking about the weather is another example.

Appendix C. Grammatical Features

Table C1. Text Features

Readability scores	Readability score should reflect examinees' need to read university-level material.
Amount of information	Although this is not a grammatical property, strictly speaking, amount of information is conveniently measured by the same tools used to measure readability. Simple measures of this characteristic may be: number of words, number of sentences, or number of clauses. Texts with high information counts increase the difficulty of even simple locate tasks.
Sentence types (simple, compound, complex, compound-complex)	The distribution of these types should be typical of university texts.
Distribution of word classes	Of particular concern are the ratio of nouns to adjectives and the frequency of nominalizations typical of university texts.
Distribution of verb types (infinitives, actives, tenses)	The texts within a test should reflect the distributions found in university texts.
Types of subordinate clauses	Some studies suggest that second-language speakers have different problems with these structures depending on their first language.

Table C2. Vocabulary Features

Word frequency	Simple counts for comparison to existing frequency databases, such as the Francis and Kucera corpus.
Semantic characteristics	The distribution of semantic features of the words, such as the proportion of abstract and concrete words[17].
Register features	Proportion of the words that are common words used in their ordinary sense, common words used in a technical way, and technical words specific to the general topic of the text.

[17] Some of this semantic information is also picked up in *Type of Information* in the task coding. There, however, it refers only to the information necessary to answer the question, while here it is a general characteristic of the text.

Appendix D. Rhetorical Properties

Table D1. Rhetorical Types

Definition	The purpose of a definition is to explicate the meaning of a term. A definition may be a sentence (e.g., "A phoneme is …"), but terms that are being introduced as part of a theory (e.g., introducing "phoneme" in a linguistics class) often have more elaborate definitions that may occupy up to a full chapter or an even larger section of a text and may be accompanied by schematic drawings, maps, or other illustrations. In general, a definition will locate the reference of the term in a class of objects and then describe features that distinguish it from other objects in the class. Encyclopedia articles are examples of extended definitions.
Description of an object or a mechanism	A description is intended to tell what something is and usually includes three features: a) the function or purpose, b) a list of the physical characteristics, and c) a description of the parts or components and how they articulate the object. Schematic diagrams are a common feature of such descriptions.
Classification/ partition	Classification/partition is concerned with the features or principles by which a collection of objects is organized into groups. These first-level groups may themselves be organized into second-level groups, etc. Thus, a classification may, but need not, result in a hierarchy. Examples include styles of music, language families, types of rocks. The emphasis may be on principles that will allow new objects to be assigned to their proper class. The principles can refer to inherent characteristics, to historical sequence, or to other features, or to a combination, as when inherent characteristics have a historical source. A classification hierarchy is often represented by a tree diagram.
Cause – effect	The goal of cause-effect analysis is to explain why something happened or happens. The presentation may be from cause to effect, as when the emphasis is on the consequences of something or some event. It may also be organized from effect to cause. The former is, of course, most commonly offered in a deductive framework and the latter in an inductive one. Commonly, a cause-effect analysis will be concerned with evidence to convince the reader of the plausibility of the cause-effect relationship. Discussions of conditions that enable or permit some event to occur are frequently framed in a cause-effect format.
Problem – solution	Problem-solution is closely related to cause-effect. The former is more likely to emphasize pragmatic connections (what you should do when X occurs), however, while the latter is more concerned with theoretical ties between the cause and the effect (why X occurs). In some cases, the problem-solution text may consist of a list of problems and a parallel list of solutions. The plausibility of the solution is often defended by its practicality. While cause-effect relations may start with either the cause or the effect, problem-solution texts almost always start with the problem.

Illustration	The presentation of examples is used to further explicate a concept. Generally, the assumption is that the examples are better known than the concept, so existing knowledge can be a connection to new knowledge.
Comparison / contrast	Formally, comparison refers to the analysis of similarities among a group of objects, and contrast, to an analysis of the differences. Nonetheless, the two are similar in that they both require the isolation of key features and the demonstration that these features have similar or dissimilar characteristics in the objects being analyzed.
Analysis	Analysis is concerned with the application of principles to new cases. It may be used to further explicate those principles or to demonstrate greater generality for them. A particularly strong form of analysis is when cases that were previously difficult to understand are shown to be easy to understand with the new principles.
Simple exposition	Some texts may be primarily lists (whether in prose or document form) with few implicit or explicit connections among the elements. A list of texts for a course is an example.
Regulatory	These texts set out what must, may, and may not be done. In an academic setting, the university calendar is the most obvious regulatory text, in that it sets out what courses must be taken for a degree and conveys other rules about academic performance.

Table D2. Interaction Types (Adjacency Pairs)

Assessment – concurrence	The first party offers an assessment of something or someone to which the second party concurs by rephrasing some part of the assessment.
Invitation – acceptance	The first party issues an invitation which the second party accepts. This pair may be preceded by a pre-invitation sequence in which the first party offers a face-saving opportunity to decline the invitation (e.g., “Are you busy Friday night?”).
Compliment – Downgrade	The first party compliments the second, but the second diminishes the force of the compliment (e.g., “It was nothing.”).

Appendix E. Text Structure Properties

Table E1. Documents As Lists

Simple	Simple lists contain only a single collection of elements. A list of courses required for a BA in sociology is an example of a simple list, as is a list of sociology courses offered in the fall term when no other information about them is provided. The elements on the list may be ordered (as when the list of students in a class is organized alphabetically by last name), or unordered (as in a list of supplies to buy at the bookstore). Searches on the former are simpler than those on the latter. If an unordered list is long, it may be difficult to determine whether or not an item is on the list. On an ordered list, however, it should be possible to easily determine whether an item is on the list (if one knows the ordering principle).
Combined	Combined lists are made up of two or more simple lists in which each element in one list is paired with an element in another list. One of the lists may be taken as the primary list (indexing list). This primary list is ordered to facilitate the location of elements on it, and so that the parallel information in the other lists can be located. An elementary combined list might be a list of course numbers with the corresponding list of course names. Elements may occur more than once in one of the lists, though this seldom happens with the primary list. For example, in a list of courses and a corresponding list of professors who teach them, each professor's name is likely to appear several times. A combined list may have many component lists. For example, the typical university timetable may be made up of corresponding lists of course numbers, course names, rooms, professors, and times. Searches on the non-indexing list are more difficult, and it may be difficult to know that all relevant information has been obtained. Thus, finding out who teaches sociological methods would be straightforward in the sociology department course list, while finding all courses taught by a certain professor would be more difficult and it may not be clear when the end of the search has been reached.
Intersecting	An intersecting list consists of three (or, rarely, more) lists which are not parallel, but which intersect and form a row and column matrix. The typical intersecting list is a television schedule which consists of a list of times, a list of channels, and a list of programs. The programs occur in the cells at the intersection of a time (usually defining the columns) and channel (usually defining the rows). In academic settings, a department may prepare a table of course offerings in a matrix format with the columns representing days, the rows representing times, and the cell entries indicating the course(s) offered at a particular time on a particular day. This makes it easy for students to locate courses that do not conflict in time. In an intersecting list, the cell entries are all of a single kind (e.g., course titles, TV programs). Many statistical tables are intersecting lists. For example, a table that lists the unemployment rates for large cities is likely to have the cities as rows, particular dates as columns, and cell entries as the actual rates for various cities during that period. The table may be designed to permit contrasts across dates, as when there are several columns, each representing a different period (e.g., months, years).

Nested	In an intersecting list, sometimes the column categories (days of the week), intersect not only with the row categories (times), but also with a fourth list, such as departments in a university. Thus, there may be a hierarchy of intersecting lists—a nested list. For a true nested list, the same types of categories must be used in each of the intersecting lists. In the university example, the days of the week recur in each of the department lists. It is not necessary for every category to appear in every higher order list, however. For example, if the Geology department has classes that meet on Saturday, but the Linguistics department does not, the day category Saturday would occur under Geology, but not under Linguistics.[18] As another example, the intersecting list of unemployment rates may have separate entries under each month for males and females; in this case, gender is nested under month.
Combination lists	Several types of lists can be joined into one list, as can several instances of a single type. For example, the intersecting list created by the statistical table of unemployment rates in different months for large cities may be combined with another intersecting list of month-to-month changes in the unemployment rates for those cities.

[18] Logically, Saturday could occur under Linguistics, but all the cells would be empty. Unless it is important to emphasize that the cells are blank, columns with all blank cells are often omitted for typographic reasons.

Other document texts can be decomposed into these types of lists. The following are some typical examples:

Table E2. Graphs, Schematics, Maps, Forms

Graphs	Most graphs are based on tables, so the underlying list structure is usually apparent. Since graphs are usually designed to represent how changes in one feature (one list) are related to changes in another (a second list), graphs must be at least combined lists and may be intersecting or nested.
Schematics	Schematic diagrams and process charts are also usually combined lists: a list of names and a list of pictographs representing the parts (schematic) or steps (process chart). Here, too, the parts may be numbered with a legend pairing the part/step name with each number. Where the schematic serves as an assembly diagram, the legend may also contain a list that provides the count of each part needed to complete the assembly. Where the whole document contains a sequence of steps with a diagram for each, the schematics may themselves be entries in a list.
Simple maps	Simple maps are combined lists. For example, in a campus map with buildings labeled, one list consists of the building names and the other of small pictographs of buildings. Often, names are not marked directly onto the map; rather, the buildings on the map are numbered and a legend pairs these numbers with the names. In this case, there are three combined lists: names, numbers, and pictographs.
Complex maps	Many maps have guides in the margins to help users locate particular sites. These marginal guides form a matrix, and the map becomes an intersecting list, with the columns and rows (typically one is a sequence of numbers and the other a sequence of letters) serving as the index to the name/place combined list that occupies each cell.
Forms	Forms are an example of complex documents. Many forms are simply a list of one-item lists. The label of a blank (such as Name) may be considered the label of the list and the entry a respondent fills in (such as Dorothy Chalker) as the single item in that list. The other blanks (address, city, etc.) also form one-item lists. Some forms are multiple-entry lists and may include a combined list. An order form has this structure: a list of quantities, a list of item descriptions, and a list of prices are all combined. Forms differ from other documents in that they require the reader to supply some of the information.

Table E3. Prose Text Structures

Records / reports	In many ways, these are the prose equivalent of simple lists. In many records, the only relationship among the various parts of the text is that they are of the same type. In other cases, the parts are of the same type and in some simple sequence. A typical record would be the notes that a student makes while observing a phenomenon. The parts are sequenced by the time of observation, but this sequence may not be related to the relationship among the observations. Another example would be notes made while reading various texts on a particular topic. In Mosenthal's scheme, reports differ from records only in that they are couched in the past tense rather than the present. The topic of the text may not be stated, and the relation between the elements of the record and the topic are mostly "an instance of'' relations; there are no explicit markers of this relation.
Generalized records / reports	In generalized records/reports, the relation of the parts to the topic is one of typicality. Such texts represent a reordering of the information in a record so that similar parts in the record are summarized. The relationship among these parts, however, is much the same as it is in a simple record; the parts go together because they relate to the topic, not to each other. Encyclopedia articles often are in the form of generalized records.
Loose classifications	These simple classifications add relationships among the lower order parts. In addition to creating similarity groups, the text may contrast and compare the resulting groups. Note that the intra-group relationship is still based on similarity relations (similar to/different from). Examples include a text that presents information on the classification of different species, a linguistics report on changes from Middle English to Modern English (primarily a contrast relationship), and a linguistics report on the features shared by Frisian and English (primarily a comparison relationship). (In the Middle/Modern English case, a loose classification would not attempt to explain why the changes occurred.) These three prose structures could easily be transformed into document structures. The Middle/Modern English contrast might, for example, be arrayed in a combined list. These three structures also focus solely on what is and seldom argue what ought to be or what naturally is.
Strong classifications	These structures are similar to loose classifications, but the inter-part relations are more complex. Here, relations of cause often occur, and the parts and their classification may serve as evidence for some claim. Alternatively, the parts and their relations may function as illustrations of a general principle, as when the changes from Middle to Modern English are shown to be consistent with standard explanations of how languages can change.

Speculatives	These structures are also rooted in cause-effect (or problem-solution) relationships, but rather than serving largely as explications of a principle, the parts and their relation are used to call into question received understandings (general principles) and to suggest and argue for a new understanding (new general principles) of the relationship. In our linguistic example, a speculative text would argue that existing principles provide a poor account of the changes and suggest that some new, different principle might provide a better account. That is, it would raise a hypothesis.
Theoreticals	These texts carry out the test implied in the hypotheses in speculative texts. In an important sense, theoreticals compare and contrast relationships rather than features in that they attempt to identify the best set of principles that are compatible with the evidence. In the English changes example, a theoretical text might show that more of the differences are viewed as natural developments under the new principles than under the old, or that the new principles provide a more economical account of the changes. A more complex theoretical text might argue that while the new principles do not provide a better account, they provide no worse one, and they do provide a better account of other linguistic changes.

Table E4. Components of Interaction

Turn-taking	A speaker's turn must start promptly at the end of the previous speaker's turn. The end of a turn is signaled by syntactic cues (it is at the end of a syntactic unit) and by prosodic cues (the speaker's intonation pattern marks a turn end, as does the direction of the speaker's gaze).
Topic	A next turn must be related to the topic of the previous turn, or the start of a new topic must be announced.
Function	The function of a turn must be congruent with the function of the previous turn. Conversational analysis captures this in the notion of adjacency pairs (see above).

Appendix F. Description of the "Type of Information" Variable

Test questions that ask examinees to identify a person, group, animal, place (as a noun), or thing are usually highly concrete—that is, these entities can easily be visualized. Hence, such tasks tend to be comparatively easy. On a scale of 1 to 5, where 1 is the easiest, questions requesting these types of information are scored 1 for difficulty. Examples of these types of questions are:

- Who invented the laser? (person)
- What animal did the Syracuse Zoo recently add to its collection? (animal)
- What is the capital of Mexico? (place)
- What building material is used to prevent the transfer of heat? (thing)

Questions that request information about amount, time, attribute, type (or kind), action, location, group, or procedure are slightly more abstract , and hence are somewhat more difficult to process, on average. On the aforementioned 1 to 5 scale, questions requesting these types of information are scored 2 for difficulty. The following examples illustrate these types of questions:

- What is the current prime interest rate? (amount)
- When was Thomas Edison born? (time)
- What color was the White House before the War of 1912? (attribute)
- What are two kinds of elephants? (type)
- What did Hamlet do after Ophelia died? (action)
- Where is the Euphrates River? (location)
- Which group of Native Americans occupied Central New York in the early 1700s? (group)
- What are the steps for making jello? (procedure)

In particular, note that an attribute is information that can qualify a person, group, animal, place, thing, or action, or even another attribute (or, traditionally speaking, another adjective or adverb).

Questions requesting information about manner, goal, purpose (or function), alternative, attempt, condition, sequence, pronominal reference, verification, predicate adjective, assertion, and problem tend to be even more difficult. On a scale of 1 to 5, such questions are scored 3 for difficulty.

Manner questions seek adverbial information about the qualification of an action or attribute, such as "How fast was the car driving?" *Goal* refers to a desired outcome (e.g., "Why did William Thomas want to become President of the AFL-CIO?"), while *purpose* refers to an intended effect (e.g., "What was the author's purpose in writing this poem?"). *Alternative* refers to a choice among two or more options ("Which alternative for reducing environmental pollution did the author advocate?").

Attempt is a traditional story-grammar category that includes actions in which characters engage to accomplish a goal (e.g., "What did Harriet Storr do to become the President of Miles Technologies, Inc.?"). (Note that "attempt" is different from "action" in that the latter describes an event that is not clearly directed towards the attainment of a goal, while the former is.) *Condition* refers to the specific states under which certain actions are prescribed. For example, in the statement, "Open the liquid content at room temperature," "at room temperature" would be a condition. Condition also refers to states that typically co-occur with actions but are not considered necessarily causal. For example, in the statement, "The plant seed grew when the soil became wet," "when the soil became wet" is a condition. *Sequence* refers to the order in which an action, attempt, or procedure occurs (e.g., "Place the following five steps for making jello in order").

Pronominal reference involves the identification of an antecedent of a pronoun (e.g., "Who does 'he' refer to in 'John tripped Jim; then he tripped Mary.'?"). *Verification* refers to whether a statement is true or false (e.g., "Is it true that Marco Polo made two separate trips to China?"). *Predicate adjective* (and predicate nominative) information refers to an attribute, person, place, animal, or thing used after the verb form "to be." For example, in the statement, "Jim Walsh is a State Senator from New York," "State Senator" is a predicate adjective. A question related to this statement requesting such information would be, "Who is Jim Walsh?" *Assertion* refers to a forceful claim for which no evidence is provided ("The man in handcuffs asserted that he had no part in the murder of his rich uncle"). *Problem* refers to a condition which blocks the attainment of a goal or the maintenance of civil operating procedures ("The problem with gangs is that their violence results in the death of innocent people"). Finally, *solution* refers to an action or procedure which eliminates a problem so that a goal can be obtained or civil operating procedures restored (e.g., "One solution to gang violence is requiring gang members to perform community service in a pediatric hospital").

Questions requesting information about the identification of cause, effect (or outcome, result), evidence (or justification), similarity, pattern, opinion, and explanation are even more abstract and difficult, on average. On the 1 to 5 scale, such questions are scored 4 for difficulty. *Cause* refers to information that produces a change in state resulting in a new state called an *effect* ("What caused the ceiling to buckle and the roof to collapse?"). *Evidence* is information that justifies a claim ("What evidence does the author provide to suggest that the earth's ozone layer has been irreparably damaged?"). *Similarity* tends to be any type of information that involves shared features or characteristics ("How are African elephants similar to Asian elephants in terms of their external physical characteristics?"). *Opinion* refers to information representing the belief or perspective of a character in terms of what is or what ought to be ("Based on the text, what is the author's position regarding logging and the Spotted Owl?"). *Explanation* consists of the enumeration of causes or reasons associated with an identifiable effect, outcome, or condition ("Explain why Joseph Conrad used the word 'shoe' so many times in his story 'The Heart of Darkness'").

Questions that require examinees to identify equivalents, differences, and themes are even more abstract and difficult, on average. On a scale of 1 to 5, such questions are scored 5 for

difficulty. *Equivalence* refers to information related to the meaning of a highly unfamiliar word or phrase. In tests and assessments, equivalence questions typically require readers or listeners to define a low frequency word or unfamiliar phrase for which no contextual clues are provided in a stimulus (e.g., "Define 'lugubrious' in the sentence, 'He felt particularly lugubrious'"). *Difference* tasks tend to involve distinctive or contrastive features related to states, events, processes, or procedures (e.g., "What are the differences between the old and new ways that American Express processes its credit-card forms?"). *Theme* includes a title (or main idea) that characterizes the most salient information in a text, or a descriptive summary of this information. In addition to questions about equivalents, differences, and themes, questions in which there is no indication of the type of information being requested also tend to be quite difficult. This type of requested information (called *indeterminate*) is often found in multiple-choice tasks in which the choices themselves represent different types of information (e.g., cause, attribute, or manner).

Appendix G. Description of the "Type of Match" Variable

The *type of match* variable refers to the processes used to relate information in a question or directive to corresponding information in a text, and to the processes used to select an answer from a range of response options. Type of match consists of a range of strategies which vary in difficulty and a variety of conditions which render processing strategies either more or less difficult (Kirsch & Mosenthal, 1990a, 1995; Mosenthal, 1996).

In *locate tasks*, respondents match one or more features in a question to one or more features in a text (Kirsch & Mosenthal, 1992b; Mosenthal, 1996; Mosenthal & Kirsch, 1991d). Based on this match, respondents then locate the answer in a sentence or paragraph associated with these features. An example task based on the text shown in Figure G1 would be: "How many job cuts did Sears announce?" To answer, readers and listeners must match the information given in the question—that is, Sears did announce job cuts—to the corresponding "Sears announced . . . job cuts . . ." in the article. Once they make this match, respondents can identify the requested information (an amount) as "50,000" in the text.

In *cycle tasks*, examinees perform an iterative series of locate searches (Mosenthal & Kirsch, 1992a; Mosenthal, 1996). These tasks may involve selecting information that meets a particular criterion or condition. The relative difficulty of cycle tasks depends on whether they draw on information within a paragraph or between paragraphs, the latter being more difficult than the former. For example, a cycle task applied to the text in Figure G1 would be: "According to the article, what are three types of small businesses that were likely to experience job growth in 1993?" Respondents must cycle to different paragraphs in the text to produce the answer: "high-tech companies, home-health-care providers, and office supply distributors."

Small Victories Make Up for Big Layoffs

Where the jobs are: Smaller companies, temporary positions and service industries.

NEW YORK (AP) -- Thousands of job cuts announced by Sears, IBM, and big aerospace manufacturers last month obscured a subtle counter-trend: smaller companies are hiring.

Economists and labor forecasters say a broad array of small to medium-sized companies, the engine of job growth during the 1980s, are at least looking to fatten their staffs as it becomes clearer that the economy is improving.

Many of these companies, which range in size from a couple of dozen workers to 1,000, already have added a few people here and there. But the effect of this marginal job creation is hard to detect and has gone largely unnoticed.

"When IBM lays off thousands of people, that more than compensates for the small companies adding 10 jobs at a time," said Andrew Campbell, president of Corporate Technology Information Services Inc., a high-tech industry research firm in Woburn, Mass., that tracks hiring plans for 35,000 companies. Sears announced 50,000 job cuts last month, big aerospace manufacturers said they'd slash 36,000 positions and IBM planned to trim 3,000.

But Campbell said his firm has seen small high-tech companies increasingly eager to recruit. In a survey completed last month, it found that nationally, these companies are planning to expand their staffs by 6.3 percent in 1993. That's about 139,000 new jobs.

Others foresee modest increases of hiring in other small businesses, ranging from home-health care providers to office supply distributors.

On the other hand, the Fortune 500 companies, which once employed 20 percent of the U. S. work force, are now down to about the 10 percent level and still shrinking, said Richard Belous, a labor economist with the National Planning Association, a research group in Washington.

"The small companies are where the job growth is going to come from," he said. "In terms of quantity, that's good news."

In addition, a growing number of companies are looking to hire temporary workers, at least, as their business improves. Manpower Inc., the nation's leading temporary help company, is likely to increase its payroll this year, breaking its record 550,000 workers in 1992.

But most job experts agree that practically all new employment will be in services, not manufacturing, where the number of workers has eroded as U. S. factories have moved operations abroad or learned to produce just as much with less help. In general, service jobs offer lower pay and fewer benefits.

The question of job creation is important because it is critical to the economic recovery, which is now technically in its 22nd month. Tuesday, the government's chief economic forecasting gauge gave some of the strongest signals yet that the recovery will last through much of 1993.

But many economists are perplexed because labor demand has been so sluggish. Historically, job creation has surged this far into an economic recovery, resulting in increased incomes and property.

Instead, the unemployment rate has remained stuck at 7.3 percent, job growth is slow and big-time layoffs are still front-page news.

Figure G1. A text used to illustrate the "type of match" and "plausibility of distractors" variables

Integrate tasks involve two steps. First, respondents must apply one or more cycle strategies to identify two or more pieces of information in a text using categories specified in the question (Kirsch & Mosenthal, 1992/1993). Then, they must relate the different pieces of information according to some type of relation in the question. This relation might be a similarity (i.e., comparison), difference (i.e., contrast), degree (e.g., smaller or larger), cause-effect relation, problem-solution relation, class and case relation, hypothesis-evidence relation, information saliency (e.g., distinguishing more important from less important information), or assertion-reason relation. In general, integrate tasks that require readers to compare information are easier than those that require them to contrast information (Kirsch & Mosenthal, 1990a; Mosenthal, 1996; Mosenthal & Kirsch, 1993b). An example of an integrate task based on the article in Figure G1 would be, "Identify three differences between service-industry vs. manufacturing jobs" (answers: "Service jobs offer low pay, fewer benefits, and are more easily filled with temporary workers").

In performing integrate tasks, respondents draw on information categories provided in the question to locate the corresponding categories in a text. They then relate the text information associated with these different categories based on a relation specified in the question (cf., Mosenthal, 1996; Mosenthal & Kirsch, 1993b). Thus, in the previous example, respondents were given the categories "service-industry jobs" and "manufacturing jobs" in the question as well as the relation "contrast."

In some cases, however, respondents must infer the categories to be searched on, or how the categories in a text relate to one another, or both, before using an integrate strategy. When such inferencing is required, respondents are said to use a *generate strategy* (Mosenthal & Kirsch, 1993a). An example of a question requiring a generate strategy would be, "Discuss whether or not the title of the article in this figure represents a good summary of the article's main point." In this case, respondents must use a generate strategy to determine that the appropriate categories to be integrated include "small company hirings" and "large company firings." Respondents must then integrate information across the article to determine that, while small companies are hiring small numbers of new workers, large companies are laying off large numbers of workers. Here, respondents must understand that what is said in the article contrasts with what is stated in the title—in short, "small victories do not make up for big layoffs." Thus, the article's title is not a good summary of the text's main point.

On average, locate tasks (scored 1 for difficulty) are easier than cycle tasks (scored 2), which in turn are easier than integrate tasks (scored 3), which are easier than generate tasks (scored 4). This is logical, of course, because cycle tasks presuppose the ability to perform multiple locate tasks, integrate tasks presuppose the ability to perform cycle tasks, and generate tasks presuppose the ability to perform integrate tasks (Mosenthal, 1996).

Additional Processing Conditions

In addition to these strategies, type of match between a question and a text is influenced by *processing conditions* that may contribute to a task's difficulty (Kirsch & Mosenthal, 1990a,

1995; Mosenthal, 1996). These conditions are identified in Figure G2. The first is the number of phrases to search on. This condition acknowledges that, as the amount of text-related information specified in a question increases, question difficulty is also increased. (Note that this does not include information that specifically describes how examinees are to respond to a question, e.g., "Find the word it, click on its antecedent, and drag the antecedent to the box below.") For instance, a question containing only one independent clause is, on average, easier than a question containing one independent clause and one dependent clause, which in turn is easier than a question consisting of one independent clause and two dependent clauses. For example, the question "How many job cuts did Sears announce?" could be made more difficult by rewording it as, "How many jobs cuts did Sears announce when making this announcement last month?".

Matching difficulty depends on the number of responses required and whether or not the number of responses, if greater than one, is specified in the question (Kirsch & Mosenthal, 1990a, 1995; Mosenthal, 1996; Mosenthal & Kirsch, 1993b). Questions requiring readers to list only one answer are easier than those requiring two or three answers, which are easier than those requiring four answers. Moreover, questions that specify the number of responses to be listed are easier than those that do not. Hence, the question, "What is one difference between service-industry vs. manufacturing jobs?" (which requires one prompted response) would be easier to answer than the question, "What are four differences between service-industry vs. manufacturing jobs?" (which requires three prompted responses). This, in turn, would be easier to answer than the question, "What are the differences between service-industry vs. manufacturing jobs?" (which requires three unprompted responses).

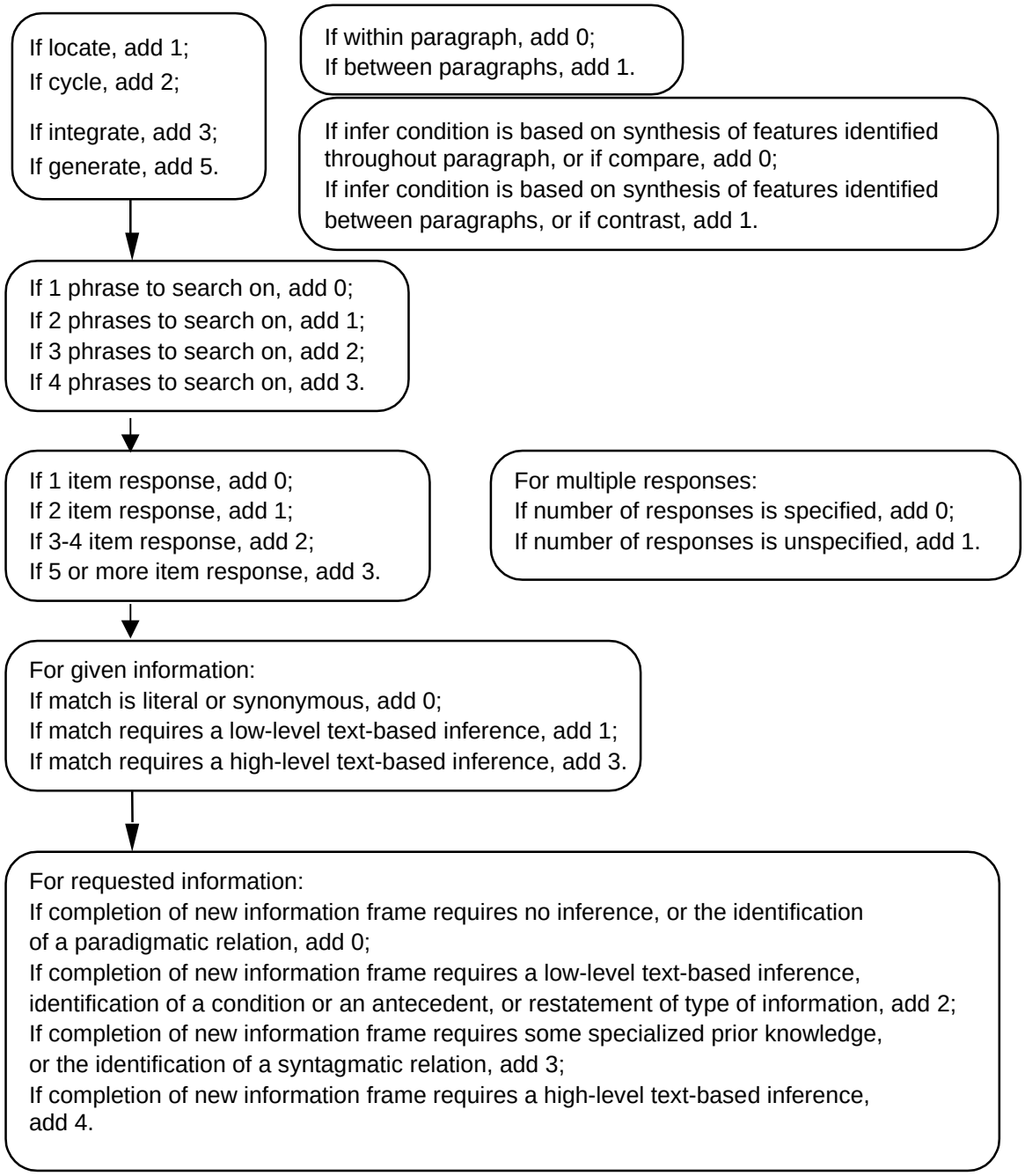

Figure G2. The additive scoring rubric used to characterize the difficulty of processing strategies and related conditions of use

Matching difficulty is also increased when readers have to make inferences to: (a) match given information in a question to corresponding information in a text and (b) identify the correct requested information (Kirsch & Mosenthal, 1990a, 1995; Mosenthal, 1996; Mosenthal & Kirsch, 1993b). In terms of (a), tasks are easier to the extent that the matches between given information in a question and corresponding information in a text are synonymous or have an easily identifiable categorical relation (e.g., "A robin is a bird."). Tasks are harder when the corresponding information in a text is not synonymous or does not have an easily identifiable categorical relation.

In terms of (b), tasks are easier to the extent that cohesive markers (e.g., because, hence, in sum) are present to signal more abstract types of information requested (e.g., manner, result, evidence, cause, difference, and theme). Tasks are more difficult to the extent that the type of information requested has to be inferred because no cohesive markers are present. Task difficulty is also increased to the extent that, in order to identify requested information, respondents must match an antecedent (usually a pronoun) with its referent (usually the referent for which a pronoun stands). In some cases, respondents must bring specialized knowledge in order to recognize text information as belonging to a particular category of information. In still other cases, completing the requested information may be more difficult if respondents must identify a condition identified outside of the node in which the answer occurs. Finally, completion of requested information is relatively easy when the identification of a word or phrase's synonym in adjacent text involves "paradigmatic context" (i.e., a context in which the synonymous word appears as the same part of speech, in the same syntactic relation, and in the context of one or more identical words as the word or phrase that it is synonymous to). Conversely, completion of requested information is made more difficult when respondents must identify a word or phrase's synonym in adjacent text involving "syntagmatic context" (i.e., a context in which the synonymous word or phrase does not meet all three paradigmatic context conditions).

Appendix H. Description of the "Plausibility of Distractors" Variable

A third variable shown to contribute to task difficulty is *plausibility of distractors* (Kirsch & Mosenthal, 1990a, 1995; Mosenthal, 1996; Mosenthal & Kirsch, 1993b). This variable has to do whether or not features of a question's given and/or requested information appear in the text but, once matched or identified, do not yield the correct information.

In general, tasks are easiest to process when there are no plausible distractors in a text. Such tasks are assigned a score of 1 for the plausibility of distractors variable. This is often the case when there is no other information in the text that meets any of the task conditions, including type of information requested. Using the text shown in Figure G1, an example of a question with no distractors would be, "According to the article, what type of victories make up for big layoffs?" (the answer is "small"). Note that, in this case, there is no mention of "victories" in the article except in the title, and only the qualifier "small" precedes "victories."

Open-ended tasks become slightly more difficult when plausible distractors for either given or requested information, but not both, appear in the text. Such tasks are assigned a score of 2 for this variable. An example: "What is the effect on the economy when IBM lays off thousands of workers?" The answer appears in the fourth paragraph of the article in Figure G1: "This more than compensates for small companies adding 10 jobs at a time" or "This offsets the growth achieved by small companies." No effect is mentioned earlier in the article, but "IBM" in the first paragraph is a distractor for the given information. A similarly difficult type of distractor occurs in multiple-choice tasks in which the question stem does not identify any given information and in which only one of the distractors contains information found in the text.

Open-ended tasks are more difficult when the text contains plausible distractors for both given and requested information, one of which may be in the paragraph in which the answer occurs. Such tasks are assigned a score of 3 for this variable. An example of such a task: "According to Andrew Campbell, what type of small companies appear to be increasingly eager to recruit new workers?" The answer appears in paragraph five, which states, "But Campbell said his firm has seen small high-tech companies increasingly eager to recruit." Note that "Andrew Campbell," a distractor for given information, is also mentioned in an earlier paragraph, and that several other types of small companies are mentioned throughout the article.

A similarly difficult type of distractor occurs in multiple-choice tasks in which the question stem does not identify any given information and in which two or more of the distractors contain information found in the text. Another type of distractor of comparable difficulty occurs in tasks requiring respondents to identify a referent to a pronoun. The referents in the distractors may share the same syntactic form as the pronoun, and are either singular or plural like the pronoun, but do not semantically complete the phrase in which the pronoun occurs.

Open-ended tasks increase in difficulty when plausible distractors for given and requested information appear in the same paragraph, but are not in the paragraph in which the answer appears. Such tasks are assigned a score of 4 for this variable. An example of this level of

distractor would be, “What type of company in the early 1990s is likely to expand its staff, breaking a record level of jobs?” The answer (“a temporary help company—Manpower Inc.”) appears in the fifth to the last paragraph in the article. However, the sixth paragraph also mentions expansion and the 1990s (related to given information) as well as type of company (related to requested information).

Open-ended tasks are most difficult when plausible distractors for given and requested information both appear in the same paragraph as the answer. (Such tasks are scored “5” for this variable.) An example task: “When this article was published in February, 1993, what percentage of the work force was employed by Fortune 500 companies?” Note that in the seventh paragraph, two percentages (representing plausible requested information) both apply to the work force employed by Fortune 500 companies (given information). The answer is “10 percent,” but “20 percent” appears in the same paragraph and would be an excellent distractor.

Note that in tasks where respondents must identify the synonym of a word or phrase in a text, a distractor given a score of 5 is one that appears in a paradigmatic context, as does the original word or phrase, but is not the correct synonym. A similarly difficult type of distractor occurs in tasks requiring respondents to identify a referent to a pronoun. Distractors at this level include cases in which the distractor referents share the same syntactic form, are either singular or plural like the pronoun, and semantically complete the phrase in which the pronoun occurs.

Test of English as a Foreign Language
P.O. Box 6155
Princeton, NJ 08541-6155
USA

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

Phone: 609-771-7100
E-mail: toefl@ets.org
Web site: http://www.toefl.org

58713-004955 • Y40M.75 • Printed in U.S.A.
I.N. 253717