®

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 21
MAY 1986

Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference

Edited by Charles W. Stansfield

EDUCATIONAL TESTING SERVICE

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of over thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program and in 1973 a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to TOEFL is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English-as-a-second-language specialists from the academic community. Currently the committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. Local research may sometimes require access to TOEFL data. In such cases, the program may provide this data following approval by the Research Committee. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1984-85) members of the TOEFL Research Committee include the following:

Henry F. Holtzclaw, Jr. (chair)	University of Nebraska
Kathleen M. Bailey	Monterey Institute of International Studies
Paul J. Angelis	Southern Illinois University
Alison d' Anglejan-Chatillon	University of Montreal
Russell N. Campbell	University of California at Los Angeles
John Haskell	Temple University-Japan

TOWARD COMMUNICATIVE COMPETENCE TESTING: PROCEEDINGS OF THE SECOND TOEFL INVITATIONAL CONFERENCE

Edited by Charles W. Stansfield

Educational Testing Services
Princeton, New Jersey

The papers in this volume were developed under a contract with Educational Testing Service. However, the content and opinions expressed do not necessarily reflect the position or policy of Educational Testing Service or the TOEFL program and no official endorsement should be inferred.

DEDICATION

This volume is dedicated to Ruth Miller, senior editor in the School and Higher Education Programs division at Educational Testing Service. For twenty-five years, Ruth assisted test program staff in the preparation of ETS reports and publications. This volume represents one of the last projects she worked on before retiring at the end of 1985. Two months later, she succumbed to illness.

CONTENTS

CONTENTS (continued)

ACKNOWLEDGEMENTS

I have always felt a little uncomfortable listing myself as editor of a report. It is not that serving as editor is a simple task. Indeed, it is a far more demanding task than most people imagine. Perhaps only those individuals who have edited volumes themselves appreciate the amount of work involved. What bothers me about being an editor is the fact that the editor gets credit for bringing a work to fruition, while the contributions of others who played a role in the the project either go unnoticed or are quickly forgotten. Therefore, it is appropriate to recognize the contributions of others who contributed to this project.

This volume is the report on a conference that was held at Educational Testing Service in October 1984. In coordinating this conference for the TOEFL Research Committee and Committee of Examiners, I was assisted by many individuals. Russell Webster, TOEFL Program Director, worked with me in planning the conference. Vera Jones made arrangements with the Henry Chauncey Conference Center at ETS. The staff of the Conference Center ably attended to the needs of the participants. Donna Natriello, my secretary, also provided assistance with these arrangements and in maintaining communication with the presenters as they prepared their papers. Atty Van Hamel tape-recorded the discussions.

In the preparation of these proceedings I was assisted by Ruth Miller and Nancy Parr, who reviewed and edited different drafts of the final report. Brenda Mahan and Karen Copper entered the original papers on a word processor, and then subsequently made hundreds of changes on the discs. Barbara Mathews handled arrangements for the printing of the volume with the ETS Publications division.

I am indebted to these individuals for their contribution to this report.

Charles W. Stansfield

INTRODUCTION

The First TOEFL Invitational Conference was an informal meeting held in Jackson Hole, Wyoming, in 1977, at the request of the TOEFL Committee of Examiners. It brought together several people active in the fields of English as a second language and language testing for a day of discussion regarding the TOEFL. Several invitees subsequently served on the TOEFL Committee of Examiners.

During the fall of 1981, 1982, and 1983, the TOEFL Research Committee and the Committee of Examiners held a joint half-day meeting in conjunction with the fall meeting of each committee in Princeton. During these sessions, and during separate meetings of each committee, the subject of communicative competence testing arose. Although committee members were familiar with innovations in the field, they wanted additional input from persons actively involved in this trend. Such input, it was hoped, would assist them in applying this subject to the TOEFL. As a result, in November 1983 the committees asked the TOEFL Policy Council to fund a Second TOEFL Invitational Conference that would focus on the subject of communicative competence testing in the TOEFL context. The proposal was approved by the TOEFL Policy Council.

During their joint meetings, the committees spent considerable time discussing the issues to be considered at the conference. Based on background papers on communicative competence prepared by Richard Duran and Charles Stansfield, the committees decided to invite five consultants to develop major papers on the subject. While there would undoubtedly be some overlap, each paper would be directed toward a different aspect of the subject. It was decided that two of the papers should focus on the implications of communicative competence for the TOEFL program, one should assess the current TOEFL from the standpoint of its effectiveness in tapping communicative competence, and the final two papers should focus on ways in which communicative competence could be measured in a modified TOEFL. In addition to selecting the five major presenters, the committees asked four other individuals to prepare formal responses to selected papers.

In November 1983, following approval by the Policy Council, the presenters and respondents were contacted. They were asked to prepare an outline of his or her papers and present it to ETS staff and representatives of both committees at a meeting held in March 1984 during the Eighteenth Annual Teachers of English to Speakers of English to Other Languages (TESOL) Convention in Houston, Texas. Committee members and TOEFL staff present at the Houston meeting reviewed the outlines with the presenters and provided them with feedback and additional information about the TOEFL program. Not all presenters attended the Houston meeting however, and the outlines of those not attending were distributed to committee members for comment by mail. All outlines were shared with all presenters, who were encouraged to communicate with each other in carrying out their assignments.

Presenters were encouraged to submit their papers for review by ETS staff and members of the committees well in advance of the Invitational Conference, thereby allowing time for revision prior to the conference. They were also requested to send their final papers to ETS one month prior to the conference for distribution to the respondents and members of the committees. All but one of the presenters distributed their papers in advance.

Presenters were sent the background papers on communicative competence that had been prepared by ETS staff for the two TOEFL committees. They were also sent an interim draft of TOEFL Research Report 17 (Duran, Canale, Penfield, Stansfield, & Liskin-Gasparro, 1985), which is a content validity study of the communicative characteristics of the TOEFL. This study had been commissioned earlier by the TOEFL Research Committee.

The Second TOEFL Invitational Conference was held October 19-20, 1984, at the Henry Chauncey Conference Center on the ETS campus in Princeton, New Jersey. (The agenda for the conference follows this introduction.) It was attended by the presenters and respondents, members of the TOEFL Research Committee and Committee of Examiners, the chairperson of the TOEFL Policy Council, and by TOEFL program administrators, TOEFL test development staff, and ETS researchers interested in the theme of the conference. A complete list of those attending also follows of this introduction. G. Richard Tucker, president of the Center for Applied Linguistics in Washington, DC, served as moderator.

The conference began with a welcome by Robert Altman, Vice President for School and Higher Education Programs, the division of ETS that serves as a home for the TOEFL program. Altman made reference to the ETS Standards for Quality and Fairness (ETS 1983), noting that these standards are designed to ensure that ETS tests reflect current theory, that tests are reviewed periodically through an internal audit process, and that test development committees composed of people of diverse backgrounds have input on test design. He also noted that this was the second invitational conference that the TOEFL program had sponsored, and that it is a TOEFL program policy to hold such conferences periodically to address important issues.

Russell Webster, director of the TOEFL program, spoke next, giving a basic introduction to the TOEFL program. He noted that the TOEFL is jointly sponsored by the College Board, the Graduate Record Examinations Board, and ETS. The program is governed by a policy council composed of representatives of universities as well as government and private agencies involved in international educational exchange. The Committee of Examiners is the oldest committee of the Policy Council, and the chairperson of the Committee of Examiners attends meetings of the Council. The Committee of Examiners is responsible for the test specifications, with reviewing test questions prior to their appearance on the test, and with having input on TOEFL research. The Research Committee includes three Policy Council members and two other members from the ESL field, as well as the chairperson of the Committee of Examiners. The Services Committee, composed entirely of Policy Council members, is concerned with the various services that the TOEFL program provides to examinees and score users. Webster noted that a Canadian advisory committee had been established to advise the program on how it can better serve Canadian needs.

John Haskell, chairperson of the TOEFL Committee of Examiners, described the responsibilities of that group. The committee is charged with the development and evolution of the test's specifications. Whenever possible, it reviews test questions prior to their appearance on the test. It also works with the Research Committee to conduct research on how the TOEFL can be improved. This conference was an outgrowth of the latter concern. The Committee of Examiners is called on to speak about the TOEFL at conferences and meetings, and to receive input on the test from professionals in the field. Haskell added that the committee looked to this conference to provide ideas about what needs to be done and how to stimulate the implementation of changes that will orient the TOEFL toward communicative competence testing.

Protase Woodford, head of the Languages Group in the test development area that assembles the TOEFL, reviewed the format of the current three-section TOEFL. He noted that the TOEFL is assembled according to a set of specifications designed by the Committee of Examiners. The items used on the TOEFL are created primarily by item writers who are practicing teachers of ESL. Currently, there are 44 item writers who work with TOEFL test development staff. These item writers are trained by ETS staff at item writing workshops that are held periodically in different parts of the country. After items are written, they are field-tested by being included on the seven long forms of TOEFL administered annually in North America. Thus, Woodford noted, the North American TOEFL is usually 50 items longer than the TOEFL given abroad in order to gather statistical data on items before they become operational.

The remaining presentations are included in the subsequent chapters of this volume. The papers by Savignon, Candlin, Bachman, Douglas, and Oller, and the reactions by Larsen-Freeman, Stansfield, and Hinofotis, are edited versions of the papers they prepared for the conference. The other papers included herein came to fruition in a different manner.

The nature of the edited documents included in this volume depends on the formality of their original format. Those that were presented in a more formal manner are included nearly verbatim, except for the stylistic changes an editor would ordinarily make. The background papers on TOEFL research by Hale and Holtzclaw were read by the presenters from handwritten drafts. Therefore, the organizational characteristics exhibited by these papers permit them to be included here nearly in their entirety. The integrative summary by Tucker, although delivered extemporaneously, was so well organized that it was possible to transcribe it and then edit it as if it originally had been a written text. Thus, it also is included in its entirety. The characteristics of the discourse used in the reactions by Larsen-Freeman to the Candlin paper and by Duran to the Bachman paper were more typical of the less carefully organized lecture style often used by the participants during discussion. Thus, a summary of their reactions are included herein, with a conscious attempt on the part of the editor to state each point concisely.

As can be seen from the agenda, three discussion sessions of approximately one hour each were scheduled during the two days in order to permit participants to question the speakers, and to reach concensus about issues and recommendations being considered. These sessions were recorded on cassette tapes. Although it is not possible to reproduce all the discussion, the most important points made are summarized here in a manner that is as faithful as possible to the original intent of the contributor. Whenever a contributor to the discussion cited the work of a particular author or educator, the appropriate bibliographic entry has been included in a references section at the end of the discussion. This style, as opposed to combining all references in a single section at the end of the volume, allows the reader to more rapidly locate the sources cited.

Following a conference such as this one, which focuses on a specific issue and its practical applications, it seems appropriate to inform the reader as to its effects on the test program. These have indeed been numerous. Some changes in the test have already been implemented based on ideas or recommendations made at the conference. Other are being considered, but an examination of their possible effects must precede their implementation. In some cases, formal research studies have been approved by the Research Committee.

In discussing changes in the test, it is important to recognize that the implementation of several changes was already underway at the time the conference was held. These were in response to the circulation of a draft report by Duran et al. (1985). The changes are enumerated on pages 92-93 of this volume.

Immediately following the conference, the Committee of Examiners began a two-day meeting that focused on its implications. Their first action was a decision to begin to "frame" (see Oller, this volume) the extended conversations and minitalks in Part C of the Listening Comprehension section. Each minitalk will begin with an announcement such as, "Following is a segment of a lecture that was given at the beginning of a class in geology." Each extended conversation will begin with an announcement such as "The following conversation between a man and a woman takes place after the woman has returned from a visit to a local tourist attraction." Such lead-ins will appear beginning with 1987 forms of the TOEFL. The committee also requested that ETS staff prepare lead-ins for reading comprehension passages and present them to the committee for review.

Based on Dan Douglas's paper (this volume), the committee requested that ETS staff prepare alternate versions of listening comprehension tasks for review. At the fall 1985 meeting of the committee, staff presented three versions of a minitalk and an extended conversation that had recently appeared in disclosed forms of the TOEFL. The first version was the regular version that had been used. This version contained scripted language only; that is, the conversation was read by the actors whose voices are heard on the tape. The second version was also based on the same script; however, the actors were told to make the conversation more natural by including pauses, expressions of hesitation or uncertainty, and false starts. In addition, sound effects were added to the tape so that, for example, if the conversation was supposed to take place in a restaurant, the sounds of people talking and dishes clanking were superimposed on the tape as background noise. In addition to the inclusion of background noises, in the third version of the listening tasks the actors were told only the subjects to be discussed and were given some facts in the form of notes for inclusion. Both of the alternative versions exhibited some problems. The committee was concerned that the sound effects in versions two and three might be distracting to examinees. The entirely unscripted rendition, version three, was especially problematic to record satisfactorily. The transition from topic to topic was not always smooth, and sometimes critical information was left out or presented awkwardly. As a result, tasks based on the third version had to be reenacted several times. Despite these initial concerns, ETS test development staff are now field testing, on a limited basis, listening tasks employing these different versions and will report their findings to the committee. The committee has also called for the inclusion of a greater number and variety of voices on the test in order to futher test examinees' ability to adapt to different stimuli.

Several research studies relevant to the issues raised at this conference have been commissioned recently by the TOEFL Research Committee. In response to concerns about TOEFL equating procedures raised by Lyle Bachman (see p. 84), Robert Boldt (1985) has begun a study to determine whether the performance of different groups of TOEFL examinees can be explained by different latent variables or whether a single variable underlies performance on each section of the test.

Gordon Hale (1985) has begun a study of the relationship between examinees' major fields and performance on reading comprehension passages. The TOEFL program has traditionally stated that the questions following each passage do not require familiarity with the subject. However, this position has not been tested through research. A finding that the text content and an examinee's major field interrelate in determining performance would suggest a need for either tests of English for specific purposes or a more careful balance of test content to ensure that examinees with certain majors are not unduly favored. A similar study has been proposed by William Angoff (1986). His study would examine the relationship between familiarity with American culture and performance on selected TOEFL items that employ an American context or setting.

Donald Powers (1986) proposes to conduct a separate validation of the Listening Comprehension section of TOEFL. He will determine the extent of the relationship between performance on the current Listening Comprehension section and comprehension of classroom lectures and other authentic listening tasks. The results could have implications for a revision of the specifications for the Listening Comprehension section or for the interpretation of the validity of Listening Comprehension scores.

Roy Freedle (1985) is conducting a study to determine the effect of students' native languages on the way they organize written compositions. The study will put to test Kaplan's (1972) hypothesis that the native language influences the organization of rhetoric. The study will also attempt to determine the effect on communication of any differences that are found.

Perhaps the most significant innovation that has occurred subsequent to the Second TOEFL Invitational Conference is the institution of a direct test of writing skills, the Test of Written English alluded to by Hinofotis in this volume (p. 178). Although the idea for this test originated prior to the conference, the frequent references to it as a desirable addition to the TOEFL reaffirmed research by ETS staff that supports the need for such a test. As a result, in November 1984 the TOEFL Policy Council approved a proposal (Stansfield, 1984) to begin planning for the test, and in May 1985 it approved a proposal to prepare for operational administration of the test (Stansfield & Adams Fallon, 1985). The first operational administration of the Test of Written English will take place on July 11, 1986.

The activities described above are an example of how a large, standardized testing program can respond to emerging ideas in the field. Undoubtedly more activities will follow as the TOEFL Committee of Examiners, the Research Committee, and ETS staff who work with TOEFL continue to digest the discussions that took place at the conference. Although the conference was designed to serve the needs of TOEFL, the proceedings are published here with the hope that they will be of assistance to others who are interested in the interface between communicative competence and test development, as well as those who are interested in TOEFL activities.

C. W. S.

References

Angoff, W. H. (1986). Content bias on the TOEFL (Proposal to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Boldt, R. F. (1985). Latent structure analysis of TOEFL (Proposal to the TOEFL Research Committee). 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 17). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1983). ETS Standards for Quality and Fairness. Princeton, NJ: Educational Testing Service.

Freedle, R. O. (1985). Language group differences in rhetorical writing patterns (Proposal to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Hale, G. A. (1985). The relation of TOEFL reading comprehension to students' major fields (Proposal to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Kaplan, R. B. (1972). The anatomy of rhetoric: Prologomena to a functional theory of rhetoric. Philadelphia: The Center for Curriculum Development.

Powers, D. E. (1986). Validation of the TOEFL listening comprehension section (Proposal to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Stansfield, C. W. (1984). Request for funding for writing project (Internal document submitted to the TOEFL Policy Council). Princeton, NJ: Educational Testing Service.

Stansfield, C. W. & Adams Fallon, M. (1985). TOEFL writing project: Report on planning stage and proposal for development stage (Internal document submitted to the TOEFL Policy Council). Princeton, NJ: Educational Testing Service.

AGENDA

Second TOEFL Invitational Conference
Henry Chauncey Conference Center
Educational Testing Service
Princeton, New Jersey
October 19-20, 1984

Friday, October 19

1. 8:50 - 9:30 a.m. Welcome and Conference Overview

    G. Richard Tucker (Moderator), Center for Applied Linguistics

    Robert Altman, Vice President, School and Higher Education Programs, ETS

2. 9:30 - 10:00 Description of the TOEFL Program

    Russell Webster, Director of Language Programs, ETS

3. 10:00 - 11:15 Description of the Current TOEFL

    Work of the Committee of Examiners
    John Haskell, Temple University in Japan

    Format, Specifications, and the Test Development Process
    Protase Woodford, Languages Group Head, ETS

4. 11:15 - 12:00 Review of TOEFL Research Program

    Current TOEFL Program Research
    Henry Holtzclaw, University of Nebraska

    Overview of Research Related to TOEFL
    Gordon Hale, Research Scientist, ETS

5. 1:00 - 1:20 p.m. The Meaning of Communicative Competence in Relation to the TOEFL Program (Oral Summary)

    Sandra Savignon, University of Illinois

6. 1:20 - 1:40 Explaining Communicative Competence: Limits of Testability (Oral Summary)

    Christopher Candlin, University of Lancaster

7. 1:40 - 2:15 Reactions to Savignon and Candlin

Diane Larsen-Freeman, Experiment for International Living

8. 2:15 - 5:00 Discussion of Presentations by Savignon, Candlin, and Larsen-Freeman

Saturday, October 20

9. 9:00 - 9:20 a.m. The TOEFL as a Measure of Communicative Competence (Oral Summary)

Lyle Bachman, University of Illinois

10. 9:20 - 10:10 Reactions to Bachman

Charles Stansfield, Associate Program Director, TOEFL, ETS

Richard Duran, University of California at Santa Barbara

11. 10:10 - 11:20 Discussion of Presentations by Bachman, Stansfield, and Duran

12. 11:20 - 11:40 Communication Theory and Testing: What and How (Oral Summary)

John W. Oller, Jr., University of New Mexico

13. 11:40 - 12:00 Communicative Competence and Tests of Oral Skills (Oral Summary)

Dan Douglas, Wayne State University

14. 1:00 - 1:30 p.m. Reactions to Oller and Douglas

Frances Hinofotis, National Education International

15. 1:30 - 3:00 Discussion of Presentations by Oller, Douglas, and Hinofotis

16. 3:15 - 4:00 Summaries of Participants' Reactions to Presentations

Diane Larsen-Freeman, Richard Duran, and Frances Hinofotis

17. 4:00 - 4:30 Integrative Overview of Second TOEFL Invitational Conference

G. Richard Tucker

Conference Participants

Robert Altman	Educational Testing Service
Denise Asfar	Educational Testing Service
Lyle Bachman	University of Illinois
Kathleen M. Bailey	Monterey Institute for International Studies
Isaac Bejar	Educational Testing Service
H. Douglas Brown	San Francisco State University
Russell N. Campbell	University of California at Los Angeles
Christopher Candlin	University of Lancaster
Sybil Carlson	Educational Testing Service
Susan Chyn	Educational Testing Service
Alison d'Anglejan-Chatillon	Universite de Montreal
Felicia DeVincenzi	Educational Testing Service
Dan Douglas	Wayne State University
Richard P. Duran	University of California at Santa Barbara
Roy O. Freedle	Educational Testing Service
Gordon Hale	Educational Testing Service
John Haskell	Temple University in Japan
Frances B. Hinofotis	National Education International
Henry F. Holtzclaw	University of Nebraska
Vera Jones	Educational Testing Service
Diane Larsen-Freeman	School for International Training
Harold S. Madsen	Brigham Young University
John W. Oller, Jr.	University of New Mexico
Joy M. Reid	Colorado State University
Sandra Savignon	University of Illinois
Charles W. Stansfield	Educational Testing Service
Francine Stieglitz	Boston University
Karin Steinhaus	Educational Testing Service
Barbara Suomi	Educational Testing Service
Angie Todesco	Public Service Commission of Canada
G. Richard Tucker	Center for Applied Linguistics
Russell Webster	Educational Testing Service
Protase Woodford	Educational Testing Service
Carlos A. Yorio	Lehman College, City University of New York

TOWARD COMMUNICATIVE COMPETENCE TESTING: PROCEEDINGS OF THE SECOND TOEFL INVITATIONAL CONFERENCE

CURRENT TOEFL RESEARCH

Henry Holtzclaw

The purpose of my brief talk is to summarize some of the current research activities of the TOEFL program at Educational Testing Service.

The TOEFL Research Committee is made up of six members, most of whom are specialists in English as a second language research. The current [1984] membership is H. Douglas Brown of the San Francisco State University Department of English; Kathleen M. Bailey, chair of the Division of American Language and Culture at the Monterey Institute of International Studies in Monterey, CA; Russell N. Campbell, chair of the English as a Second Language section at UCLA; Allison d'Anglejan-Chatillon of the faculty of Sciences of Education at the University of Montreal, Quebec; and John Haskell of Northeastern Illinois University, who for this year is at Temple University in Tokyo, Japan. Louis A. Arena, who is Director of the University Writing Center at the University of Delaware, has just completed a term on the Research Committee and serves for John Haskell when it is not possible for John to be at our meetings this year. Charles Stansfield, who is Associate Program Director for TOEFL, is the staff member who works with our committee, and Gordon Hale acts as liaison from the ETS Research division. It is the responsibility of the TOEFL Research Committee to provide direction for the TOEFL research program. In doing so, the committee determines appropriate topics of research for TOEFL and other TOEFL program instruments, including Pre-TOEFL, SLEP, TSE, and SPEAK. Funds for research are included in the overall budget for the TOEFL program each year and amount to approximately a half million dollars. This year, for example, we have about $575,000. Next year it should be approximately the same. To date most TOEFL research funds have been appropriated for studies of TOEFL and TSE.

A research study is funded according to the following procedure: A precis for a study is presented to the committee by an ETS staff member. If the committee approves the precis, a small proposal-writing grant is awarded from committee funds. At a subsequent meeting the researcher presents a formal proposal for the study. If the committee funds the study, the researcher begins work on the project. After the study is completed, the researcher submits a draft final report to the committee for approval. The committee may approve the report as submitted or approve it with revisions. In cases where the committee does not approve the report, a revision is submitted for approval at a later meeting.

In the remainder of my talk I will try to give you some idea of the nature of the research projects that have been completed recently, are in progress, or are being considered for future research.

First I will give some examples of research recently completed and for which the committee has recently considered drafts of the final reports. Most of these were considered at our meeting two weeks ago. One is a study on test speededness under number-right scoring that was done by Isaac Bejar (1984). A test is speeded when some portion of the test-taking population does not have sufficient time to attempt every item in the test within the allotted time. In the multiple-choice TOEFL test, the total score is based on the number of items answered correctly. On such a test, one effect of speededness could be that the students who run out of time answer the remaining items in a more or less random fashion. Clearly, to the extent that a test is speeded and students engage in random responding, responses to some items will not depend on the student's level of knowledge, which of necessity tends to reduce the validity and reliability of total scores. Bejar reached the conclusion that speed does not play a significant role in the TOEFL as a whole. A number of other conclusions were reached with respect to more general application to assessing speededness for number-right scoring procedures.

A second research study, conducted by Richard Duran (1985) and a number of associates, was on the subject of TOEFL from a communicative viewpoint on language proficiency. This study was designed to examine the content characteristics of the Test of English as a Foreign Language from a communicative perspective, based on current research in the area of applied linguistics and language assessment. The report is a lengthy and very useful one that you may have read in connection with the present conference. Another portion of Dr. Duran's work is the investigation of the item content of the three sections of the TOEFL in light of developments in the areas of teaching and assessment of communicative skills. The project in fact studies the appropriateness of the TOEFL as a test of extended communicative skills, given the present purpose of the test as a measure of basic proficiency skills in English. Suggestions are formulated about the value and possibility of new item types and the prospects of development of new instruments for communicative skills assessment.

A third research report, on the subject of unusual test behavior in the TOEFL population, was carried out by Philip Oltman (1984). Usually, as would be expected, examinees tend to make errors on more difficult items and to answer the easier items correctly. However, some unusual examinees miss some easy items and get some more difficult items correct. The aim of this study was to determine the extent to which the native languages of examinees are related to unusual response patterns on the test. One interesting observation of the report was that, compared with their relative proportion in the sample, Chinese speakers were less likely to have highly unusual response patterns while Arabic speakers were more likely to deviate from the usual response pattern.

A fourth report was on the conceptualization of a computer-delivered TOEFL for the two-year college market, this research having been done by Sybil Carlson (1984) and two colleagues. This represented a very large project and was reported in two sections representing two related studies: (1) the interpretation of data collected from a survey of two-year colleges that enroll large percentages of students for whom English is a second language (these two-year colleges are a potential market for a new instrument) and (2) a feasibility analysis of the potential for a computerized adaptive English language proficiency test as a new TOEFL program instrument. In the first study the conclusion is made that the two-year college market for a computer-delivered English proficiency test may, at this time, be too small to warrant the development of such a product targeted specifically to this market. The second report recommends that the TOEFL program consider the future development of some form of computer-delivered English language proficiency test for a larger market that would include, but not be restricted to, two-year colleges. Despite numerous risks involved, the computer delivery of tests is expected to be the delivery mode of the near future. This report is a part of an ongoing study of the possibility of computerization for other ETS tests as well as for TOEFL.

Let's turn our attention now to some typical examples of progress reports for research projects currently under way. One of these is an analysis of test-taking patterns and score change for TOEFL repeaters, by Kenneth M. Wilson (1986). The purpose of this project is to identify from TOEFL files examinees who have repeated the test one or more times and to determine the extent and nature of change in TOEFL performance as a function of such variables as time between testings, initial score level, sex, age, country of origin, native language, and the like. Some have taken the TOEFL test as many as eleven times. Important and useful data are being collected, and many resulting significant conclusions are being reached.

Another current research project is a study on writing performance of foreign students and its relationship to TOEFL scores, by Sybil Carlson and Brent Bridgeman (1985). The project investigates the relationship of scores on the current TOEFL to the kind of essay performance that would be required of beginning undergraduate and graduate students in the three fields that enroll the largest numbers of foreign students--business, engineering, and social science. Three language groups are represented--Arabic, Chinese, and Spanish--as well as native speakers of English. Two thousand six hundred and forty essays have been scored and analyzed.

A project relating to the Test of Spoken English, recently completed by Isaac Bejar (1985) aims to develop a procedure of identifying examinees who are likely to obtain discrepant ratings when rated by two examiners. It also studies whether a procedure can be validated such that it would be feasible to rate most examinees with a single rater and use a second rater only with those examinees that would be likely to obtain a discrepant second rating. I might mention that on the basis of this study the thinking is that, for the present at least, we will need to retain the system of two raters on the TSE test.

Another study currently under way is a study of academic demands related to listening skills of nonnative students, by Donald Powers (1985). The purposes of the project are (1) to develop a taxonomy of listening skills demanded of nonnative students, (2) to specify criterion tasks against which performance on the TOEFL Listening Comprehension section can be gauged, and (3) to plan a construct validity study of the listening comprehension section.

Marilyn Hicks (1984) is working on a project that is designed, through a two-stage testing procedure, to compare the results of computerized testing with paper-and-pencil testing. This project includes an attitudinal survey of computerized versus paper-and-pencil tests conducted by Charles Stansfield.

The role of cloze items in the TOEFL is the subject of much interest currently. One project that is now underway by Gordon Hale studies the possibility of a multiple-choice cloze test to provide an effective method of assessing key aspects of English language proficiency. In the cloze test studied in this project, four alternative words are provided from which the examinee must select the best one. The study of a third type of cloze test is proposed for future research and is a new research proposal that I will mention next, just approved at the meeting of the Research Committee yesterday.

The committee considered two new research proposals for which precis had been approved earlier. One of these is a study of the development of cloze-elide tests of English as a second language, by Winton Manning (1984, 1986). This project studies the use of a type of cloze testing, designated as cloze-elide testing, which involves inserting extra words into the running text that must then be eliminated by the examinee. The examinee, in reading the passage, draws a line through the extra words that must be eliminated from the passage for it to make sense. I mentioned different forms of cloze testing a bit earlier. The first of those, in which the student places a suitable word in a blank within a paragraph, is necessarily scored manually. The second, in which multiple-choice options are provided, can be machine-scored. The Cloze-Edit testing just mentioned can also be machine-scored by a unique new method developed and patented by Dr. Manning. Particularly the latter two types represent possible new testing techniques for future TOEFL tests.

Another new proposal for which a precis was approved and that now is under consideration by the Research Committee, is the relationship of TOEFL response choices to native language group membership and other examinee characteristics, by Philip Oltman and Lawrence Stricker (1984). Native language groups among the TOEFL population may differ from each other not only in which items they find more difficult but also in which answer options they choose when they do make errors. The purpose of this study is to determine whether different native language groups show systematic differences in the patterns of options they select on the TOEFL and whether such differences cluster together in ways that can be understood by reference to linguistic or other factors.

In conclusion, I would point out that the brief description of research projects of TOEFL completed, underway, or projected will serve to indicate the variety of projects undertaken. It is always risky and perhaps unfair to the experimenter to summarize so briefly an extensive research project. What I have hoped to do is to provide some perspective to you without doing violence to the research. Obviously, it is impossible to do more than a cursory summary in the short period of time alloted to me today. Nevertheless, I hope it has been interesting to you and helpful in gaining some insight into current TOEFL research activity.

References

Bejar, I. I. (1984). Test speededness under number-right scoring: an application to the Test of English as a Foreign Language (Final report submitted to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Bejar, I. I. (1985). A preliminary study of raters for the Test of Spoken English (TOEFL Research Report 18). Princeton, NJ: Educational Testing Service.

Carlson, S. B., Kline, R. G., & Ward, W. C. (1984). Conceptualization of a computer-delivered TOEFL for the two-year college market (Final report submitted to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

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

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 17). Princeton, NJ: Educational Testing Service.

Hicks, M. M. (1984). Development and investigation of computerized and paper-and-pencil placement tests for the TOEFL via two-stage testing procedures (Proposal submitted to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Manning, W. H. (1984). Development of cloze-elide tests of English as a second language (Proposal submitted to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Manning, W. H. (1986). Using technology to assess second language proficiency through cloze-elide tests. In C. W. Stansfield (Ed.), Technology and language testing (pp. 147-165). Washington, DC: Teachers of English to Speakers of Other Languages.

Oltman, P. K. (1984). Unusual test behavior in the TOEFL population (Final report submitted to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Oltman, P. K., & Stricker, L. J. (1984). Relationship of TOEFL response choices to native language group membership and other examinee characteristics (Proposal submitted to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Powers, D. E. (1985). A survey of academic demands related to listening skills (TOEFL Research Report 20). Princeton, NJ: Educational Testing Service.

Wilson, K. E. (1986). Patterns of test taking and score change for examinees who repeat TOEFL (Final report submitted to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

AN OVERVIEW OF RESEARCH RELATED TO TOEFL

Gordon Hale

I've been asked to talk about previous research on the TOEFL and, in particular, studies relating the test to certain integrative and direct measures of English proficiency, such as the cloze and dictation tests, as some of the conferees here have called for looking at the possible role of such measures in the TOEFL. Incidentally, I use the term "direct" measure because it's a recognized label for such assessment techniques as interviews and essay tests, although Lyle Bachman (this volume) correctly points out that the term is something of a misnomer and that any performance task is really an indirect index of an underlying construct.

Charles Stansfield, Richard Duran, and I (Hale, Stansfield, & Duran, 1984) recently summarized studies involving the TOEFL--some done at ETS, but most done elsewhere. In preparing my remarks, I've focused on those studies that have dealt with the cloze test, dictation, interviews, and essay tests; and I will be concerned only with the results involving relations with the TOEFL and its subtests. I'd like to start with studies of the cloze test, since these are the most numerous. This task, as you know, is one in which the examinee reads a segment of text from which words have been deleted and is asked to indicate the deleted words.

Many people have noted that the cloze test can be a useful index of proponent of this method (e.g., Oller, 1979), and others participating here hold similar views. A key research issue concerns the extent to which the TOEFL measures processes similar to those tapped by the cloze test. Also, looking at relations with the TOEFL subtests can help determine whether the cloze is more highly related to reading comprehension, to knowledge of English structure, or to other components of proficiency.

We found nine studies that used the standard completion cloze test (Darnell, 1970; Flahive, 1980; Hinofotis, 1980; Irvine, Atai, & Oller, 1974; Mullen, 1979; Pike, 1979; Ratchford, 1982; Riggs, 1982; Scholz & Scholz, 1981). In studies in which two or more different tests or passages were used, or two or more subgroups of examinees, I've taken the median of the correlations for convenience. Where both exact and acceptable word scores were computed, I've taken the correlation with acceptable word score, although usually there was little difference. The range of correlations with the TOEFL across all nine studies was .66 to .84. Where reliabilities were reported for the cloze test, they were generally in the mid 80s, so that the correlations corrected for attenuation were on the order of 10 points higher than the actual correlations. In general, then, there appeared to be a reasonable amount of overlap in the processes measured by the cloze test and the TOEFL.

Regarding correlations with the TOEFL subtests, I should first mention that all of these studies were done with the five-part TOEFL, which was used until 1976. In this version, the subtests were much like those in the current TOEFL, although further subdivided. There was a separate listening comprehension section. Two sections, English Structure and Writing Ability, were somewhat comparable to the present Structure and Written Expression section, and there were two separate sections called Reading Comprehension and Vocabulary.

Only five of the nine cloze studies reported correlations with TOEFL subtests, and, of these, just three reported reliabilities, which allowed correction for attenuation (Hinofotis, 1980; Pike, 1979; Scholz & Scholz, 1981). In these cases, corrected correlations were highest with reading comprehension--in all three cases, being in the 90s. Corrected correlations with the Writing Ability subtest were reasonably high as well, ranging from .76 to .89. So, while the cloze test appears to have the greatest affinity with Reading Comprehension, it also seems to tap processes related to recognitory aspects of writing ability.

In the two other studies in which subtest scores were examined but correction for attenuation was not possible (Darnell, 1970; Irvine, Atai, & Oller, 1974), the highest correlation, surprisingly, was with Listening Comprehension. Perhaps this was partly because the Listening Comprehension subtest of the five-part TOEFL had the highest reliability. Whatever the basis for this finding, it shows that the cloze test may tap several components of language proficiency.

Multiple-choice cloze procedures were also used in a couple of studies in which relations with TOEFL subtests were assessed (Pike, 1979; Scholz & Scholz, 1981). In a multiple-choice cloze test examinees are given several options to select for insertion into each blank. Again, the relations were highest with the Reading Comprehension and Writing Ability subtests, although relations with Vocabulary and English Structure were nearly as high, further suggesting that a cloze procedure taps various aspects of proficiency.

Another task that has been suggested for study in connection with the TOEFL is dictation, in which examinees would have to write, verbatim, a portion of text read to them. We found only one study relating dictation to the TOEFL (the five-part version), and this study got an uncorrected correlation of .69 with total TOEFL score (Irvine, Atai, & Oller, 1974). The highest relation with a TOEFL subtest was with Listening Comprehension, although the relation with English Structure was nearly as high. Thus, while dictation has been suggested as having some promise as an alternative item type, research has just begun to examine its relation to the TOEFL, and no firm conclusions can yet be drawn on this matter.

Let me turn now to studies using interviews. Here we found that, in two studies using a relatively large _N_ and a Foreign Service Institute interview (now known as the Interagency Language Roundtable, or ILR, procedure), the correlations with total TOEFL were around .70, and corrected correlations were in the low .80s (Clark & Swinton, 1979; Pike, 1979). The Test of Spoken English was used in one study (Clark & Swinton,

1980) and was found to correlate in the low .70s with the TOEFL, and near .80 when corrected. Still other studies using the Grabal and Ilyin interviews, as well as an experimental procedure, got correlations that ranged from under .50 to just over .70 (Gradman & Spolsky, 1975; Mullen, 1978). These last correlations may have been relatively low due to small Ns or to the use of more homogeneous populations rather than to the nature of the interview procedures; this can't be determined without further study. But a reasonable conclusion might be that, when an established interview technique such as the ILR procedure is used, along with a suitable sample, a moderately strong relation with the TOEFL will be found.

Regarding TOEFL subtests, two studies used the three-part TOEFL and the ILR interview (Clark & Swinton, 1979; Clark & Swinton, 1980). In both cases, corrected correlations were in the low to mid .80s for Listening Comprehension and in the low to mid 70s for the other two sections. In two studies using the five-part TOEFL and either the ILR or an experimental interview (Mullen, 1978; Pike, 1979), the corrected correlation was highest with Listening Comprehension in one case, and roughly equal for Listening Comprehension and English Structure in the other. In general, interview performance seems to be most closely linked with performance on the Listening Comprehension part of the TOEFL, which is not surprising, given that this is the one part of the test that deals with oral skills. Nevertheless, enough data show a possible link to other portions of the TOEFL to merit further study of them as well.

The issue of essay test performance and its relation to the TOEFL was examined in three studies, all using the five-part test (Osanyinbi, 1975; Pike, 1979; Pitcher & Ra, 1967). Corrected correlations with TOEFL score ranged from .70 to the low 80s, and in every case correlations were highest with the English Structure and Writing Ability sections. Thus, of the aspects of proficiency measured by the TOEFL, those concerning identification of proper structures and recognitory aspects of writing ability appear to be most highly related to direct assessment of writing.

There is one other category of studies that, theoretically, should provide reasonably direct information about students' English proficiency and its relation to the TOEFL; these are studies looking either at faculty members' ratings of students' English proficiency or the grades they give students in ESL classes. Although we did find several studies of this type, unfortunately they present problems of interpretation. For example, ratings of different students were typically made by different faculty members, so a uniform scale could not be established. And, where ESL grades were examined, the students were generally taken from several different levels of ESL instruction, again making it impossible to derive a uniform scale. In still other cases, the Ns were very small--20 or fewer. Although it would be useful to know how the TOEFL relates to direct observation of students' English proficiency--especially by someone who has interacted with the students over an extended period of time--sound research on this topic remains to be done.

If I were to try to summarize these various results, perhaps the most appropriate general observation would be that each of the integrative or direct measures is reasonably highly related to the TOEFL, suggesting a fair amount of commonality in processes measured by them. At the same time, though, the relationships are by no means perfect, implying that there may still be some aspects of proficiency tapped by these other measures that are not captured by the TOEFL.

Regarding the TOEFL subtests, the highest correlations generally tended to appear where one would expect them--for example, Listening Comprehension was the subtest most highly related to interview performance. Nevertheless, for each task, moderately strong relations were observed with two or more subtests, indicating that each of the integrative or direct measures--cloze, dictation, interview, and essay--taps more than one of the processes measured by the TOEFL.

I should point out that it was not my objective to present a review of results bearing on processes in the integrative or direct measures. As you know, there is a vast literature on these tasks, which I couldn't begin to summarize here. My goal was simply to give you an overview of the relations found between these measures and the TOEFL. Some of the papers prepared for this conference suggest that such tasks should be considered for use in connection with the TOEFL. This summary will, I hope, provide useful background information about how these tasks relate to processes measured by the TOEFL, according to the results of research done so far. I'd also like to note that there are limitations in many studies we reviewed. If further work is to be done on the role of new methods of assessment--especially as potential adjuncts to the current TOEFL--certain standards of methodology need to be met. As an example, the nature of the sample is critical. Among other things, it must be large enough to produce replicable results, and it must be heterogeneous in proficiency, so that the magnitudes of the observed relations are not artifically reduced. Of course, the measures need to be reliable enough; and to the extent that reliability is less than perfect, this fact should be taken into account--for example, through correction for attenuation--in drawing inferences about relations among the constructs being measured.

Finally, it is important to bear in mind the primary function of the TOEFL and the way in which it is administered. The TOEFL is used principally in admissions decisions, and it must be amenable to mass administration and scoring. This means that a good deal of effort would need to be put into developing types of tasks that meet the necessary restrictions. In the area of cloze testing, a group of us is currently studying the feasibility of using a multiple-choice cloze procedure and assessing its role in the factor structure of the TOEFL. Another project, being conducted by Winton Manning at ETS, is looking at the feasibility of a machine-scorable, cloze-elide test, in which examinees cross out words that have been randomly inserted in text. And the procedure Lyle Bachman describes in his chapter in this volume is another variant of the cloze that may have potential. These projects exemplify the kinds of efforts needed to develop new testing methods that could be used in the TOEFL context. If, indeed, it is desirable to consider new testing methods, one of our most challenging tasks will be to come up with techniques that allow for mass administration and scoring yet, at the same time, do not lose the essential characteristics that made those methods attractive in the first place.

References

Clark, J. L. D., & Swinton, S. S. (1979). _An exploration of speaking proficiency measures in the TOEFL context_ (TOEFL Research Report 4; ETS Research Report No. 79-8). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 201 641)

Clark, J. L. D., & Swinton, S. S. (1980). _The Test of Spoken English as a measure of communicative ability in English-medium instructional settings_. (TOEFL Research Report 7; ETS Research Report No. 80-33). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 218 960)

Darnell, D. K. (1970). Clozentropy: A procedure for testing English language proficiency of foreign students. _Speech Monographs_, _37_, 36-46. Also printed as "The development of an English language proficiency test of foreign students, using a clozentropy procedure." Final Report, U.S. Office of Education Project No. 7-H-010, 1968. (ERIC Document Reproduction Service No. ED 024 039)

Flahive, D. E. (1980). Separating the _g_ factor from reading comprehension. In J. W. Oller, Jr., & K. Perkins (Eds.), _Research in language testing_. Rowley, MA: Newbury House.

Gradman, H. L., & Spolsky, B. (1975). Reduced redundancy testing: A progress report. In R. L. Jones & B. Spolsky (Eds.), _Testing language proficiency_. Washington, DC: Center for Applied Linguistics.

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). _Summaries of studies involving the Test of English as a Foreign Language, 1963-1982_. (TOEFL Research Report 16; ETS Research Report No. 84-3). Princeton, NJ: Educational Testing Service.

Hinofotis, F. B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller, Jr., & K. Perkins (Eds.), _Research in language testing_. Rowley, MA: Newbury House.

Irvine, P., Atai, P., & Oller, J. W., Jr. (1974). Cloze, dictation and the Test of English as a Foreign Language. _Language Learning_, _24_, 245-252.

Mullen, K. A. (1978). Determining the effect of uncontrolled sources of error in a direct test of oral proficiency and the capability of the procedure to detect improvement following classroom instruction. In J. L. D. Clark (Ed.), _Direct testing of speaking proficiency: Theory and application_. Princeton, NJ: Educational Testing Service.

Mullen, K. A. (1979). More on cloze tests as tests of proficiency in English as a second language. In E. J. Briere & F. B. Hinofotis (Eds.), Concepts in language testing: Some recent studies (pp. 21-32). Washington, DC: Teachers of English to Speakers of Other Languages.

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

Osanyinbi, J. A. (1975). A concurrent validity study of the West African School Certificate and General Certificate of Education English Language Examination, using Educational Testing Service's Test of English as a Foreign Language as the criterion measure (Doctoral dissertation, University of Wisconsin, 1974). Dissertation Abstracts International, 35, 5130A-5131A. (University Microfilms No. 74-22, 134)

Pike, L. W. (1979). An evaluation of alternative item formats for testing English as a foreign language. (TOEFL Research Report 2; ETS Research Report No. 79-6.) Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 206 627)

Pitcher, B., & Ra, J. B. (1967). The relation between scores on the Test of English as a Foreign Language and ratings of actual theme writing (Statistical Report No. 67-9). Princeton, NJ: Educational Testing Service.

Ratchford, D. L. (1982). Reading ability of entering freshmen international students at a southwestern state university: Some implications (Doctoral dissertation, University of Oklahoma, 1981). Dissertation Abstracts International, 42, 3088A-3089A. (University Microfilms No. 8129435)

Riggs, J. M. (1982). Cloze testing procedures in ESL: A prediction of academic success of foreign students and a comparison with TOEFL scores (Doctoral dissertation, Indiana University, 1981). Dissertation Abstracts International, 42, 5048A. (University Microfilms No. DA 8211187)

Scholz, G. E., & Scholz, C. M. (1981). Multiple-choice cloze tests of EST discourse: An exploration. Paper presented at the fifteenth annual TESOL convention, Detroit. (ERIC Document Reproduction Service No. ED 208 656)

THE MEANING OF COMMUNICATIVE COMPETENCE IN RELATION TO THE TOEFL PROGRAM

Sandra J. Savignon

INTRODUCTION

Competence, proficiency, standards. These terms are familiar today in discussions not only of second language programs but of educational programs in general. Both learners and teachers are under scrutiny at all levels of instruction, primary grades through graduate school.

Concern for competence, or proficiency, as opposed to number of years of instruction and/or degrees earned is not new. Yet the current economic and political climate lends to voices expressing such concern a new sense of urgency. At stake, or so it seems, is not only national pride, but national survival. Faced with a documented decline in learner achievement in recent years and the flight of talented young men and women away from teaching and toward more promising careers in business, law, and other fields, many American states are currently considering or already have adopted competency requirements for teachers. In practice, such requirements typically mean the attainment of a minimum score on a test of some kind.

The field of medicine offers yet another example of the current concern for competence. Now that the historical shortage of physicians practicing in the U.S. has become a surplus and American medical schools have correspondingly reduced enrollments, there is increasing competition from physicians trained abroad. The American Medical Association recently announced that the basic test taken by graduates of foreign medical schools to gain a residency in an American hospital is being made "significantly more difficult." The test was devised by the Educational Commission on Foreign Medical Graduates, a Philadelphia group created to monitor the performance of students in medical schools outside the United States. (Lyons, 1984, p. 21)

To be sure, the rationale in each of the preceding illustrations of the current concern for competence is different, and herein no doubt lies the key to predicting the ultimate acceptance or rejection of the corresponding response. One has only to compare the annual average salary of a secondary school teacher ($21,000) with that of a physician ($110,000) to appreciate the incentive for making higher demands of the latter while tolerating mediocrity in the former. Standards of competence are relative. They depend upon what's at stake.

Inextricably tied to the question of standards is, of course, that of knowing what it is one is trying to measure in the first place. Whether the concern is with the proficiency of teachers or with that of physicians, the answer to the question of what should be tested is far from clear-cut. Harvard psychologist David McClelland has gained recognition in the business world for this efforts to define and assess competence for a particular job. His competency theory asserts that standard aptitude tests are crude measures, irrelevant, in the main, to real-life job success. He has shown the incongruity between the kinds of aptitude tests typically administered by prospective employers--standard intelligence and aptitude tests similar to those required of applicants to college and professional schools--and the job itself. These tests are undoubtedly helpful in predicting academic success, acknowledges McClelland (1973), since academic success is based in large part on more tests of a similar nature. As he puts it, "The games people are required to play on aptitude tests are similar to the games teachers require in the classroom" (p. 1). There is little evidence, however, that students who do well on the aptitude tests and earn good grades will excel in their careers. The poorer students have just as much chance of success in life as their straight-A peers, provided they have attained the same level of education and have earned the same qualifying diplomas.

Support for McClelland's view was cited in a recent New York Times report (Goleman, 1984):

> Intelligence as measured by I.Q. tests appears to be related to success in school, but it seems to have surprisingly little to do with achievement in careers, according to a growing number of psychologists. Their research has repeatedly shown that, although the best executives almost always do at least moderately well on I.Q. tests, their ranking on these tests is simply not the factor that distinguishes those who advance from those who do not. (p. 12)

The American Medical Association concurs with the need to improve the validity of competence measures. In announcing its plan to increase the difficulty of the basic test for graduates of foreign medical schools, the AMA disclosed the development of tests that would examine the ability of medical students to reason through cases that were presented to them rather than merely repeat a set of facts. Roy Schwarz, head of the association's department of education, explained that, in contrast to the one- or two-day multiple-choice tests given now, the ideal examination might last as long as a week, use actors as simulated patients, and set up computer simulations of medical problems using videotapes. "Central to the test issue is the ability of a medical student to use clinical judgment, and whether this can be accurately defined on a test." (Lyons, 1984:21) Discussions of competency tests for teachers have raised similar issues, and while there have been some efforts to implement tests of equivalent clinical skills in education--e.g., the use of an oral interview to certify foreign language teachers in Texas--budget and time constraints have more often than not dictated the use of an existing multiple-choice test. Where such is the case, expediency overrides the issue of validity.

Perhaps the most important point that McClelland makes is that the characteristics tested and the ways of improving on these characteristics should be made public and explicit. According to McClelland, this principle is in sharp contrast to the long tradition in psychometrics of keeping test items a secret. This secrecy is motivated, in large part, by the fear that if applicants have prior knowledge of the nature of the items, they will be able to practice and earn a high score--one that may no longer correlate with subsequent performance--and thereby destroy the predictive power of tests. This has resulted in another flourishing tradition: the numerous and successful commercial schools, workshops, and self-help materials that provide practice in test-like tasks for enterprising applicants. How much simpler it would be, concludes McClelland, to make explicit to the learner the criterion behavior that will be tested. The psychologist, teacher, and learner may then collaborate openly in working to improve the applicant's performance on directly observable task-related skills.

Such, then, is the general backdrop against which I would like to consider the meaning of communicative competence in relation to the TOEFL program. That the TOEFL Policy Council should sponsor a conference to consider the potential contribution of current theories of teaching and testing communicative competence to TOEFL content and format is in itself noteworthy. This and an earlier conference held in 1977 acknowledge that the validity of this widely used test of English proficiency remains an issue, an issue that those responsible for TOEFL development and administration continue to address. Specifically, we have been asked to consider to what extent a TOEFL score is an accurate reflection of a candidate's ability to use English for a wide range of academic purposes. That is, how good is the TOEFL at predicting a nonnative speaker's ability to interpret, to express, and to negotiate meaning in U.S. academic contexts? For the 450,000 or so candidates who take the test each year and whose future academic plans often depend on the outcome, the issue clearly is not to be treated as if it were one of purely theoretical interest.

I. COMMUNICATION, COMMUNITY, AND COMMUNICATIVE COMPETENCE: THEORETICAL CONSIDERATIONS

> Normally, in out-of-school conversations, our focal attention as speakers and listeners is on the meaning, the intention, of what someone is trying to say. Language forms are themselves transparent; we hear through them to the meaning intended. But teachers, over the decades if not centuries, have somehow gotten into the habit of hearing with different ears once they go through the classroom doors. Language forms assume an opaque quality. We cannot hear through them; we hear only the errors to be corrected. (Cazden, 1976:79-80)

The creative and complex nature of human communication continues to elude description and classification by those who seek to identify and to measure its components. No longer limited to the consideration of sentence-level structural linguistic features--pronunciation, and syntax--descriptive perspectives on communication have continued to broaden, taking into account a seemingly ever-increasing number of variables. In the process, linguistics, a field of study concerned with the description of language codes, has met with ethnography, a field of study concerned with the description and analysis of culture. The ethnography of communication is a synthesizing discipline that takes language first and foremost as a socially-situated cultural form. While such description is far from neat and does present what may prove to be insurmountable problems, researchers who commit themselves to such efforts do so because in their view "to accept a lesser scope for linguistic description is to risk reducing it to triviality" (Saville-Troike, 1984).

Second-language methodologists increasingly agree. Coinciding with the introduction by Hymes (1971) of the term "communicative competence" to denote the sociolinguistic features of communication, Savignon (1972) focused on the strategic competence required of learners to interpret and convey meaning in a second language, highlighting the distinction between an ability to perform well on discrete-point tests of linguistic knowledge and an ability to participate in a communicative setting involving one or more interlocutors. While the focus in both these instances was on oral communication, rich in nonlinguistic as well as linguistic cues to meaning, similar dimensions exist for reading and writing. Breen and Candlin (1980) have summed up this expanded synthetic view of communication as the interpretation, expression, and negotiation of meaning.

While the term "comunicative competence" may be new in discussions of language proficiency, the perspective it offers on language and language use has been around for some time. One has only to consider the observations of Comenius, writing in the 17th century, to understand the longstanding tension that has existed between an emphasis on language form and language function. Comenius is well known in the history of language teaching methodologies for his objection to the method of second language teaching that had resulted from the teaching of skills of grammatical analysis in the Middle Ages. The preoccupation with grammatical analysis had grown so that by the Renaissance it was viewed as a method for actually teaching the language. In the words of Comenius, "Right from the very beginning of the course, youngsters are driven to the thorny complexities of language; I mean the entanglements of grammar. It is now the accepted method of the schools to begin from the form instead of the matter, from grammar, rather than from authors..." (1648, cited in Kelly 1969:227).

The functional analysis of language has a long tradition in linguistic inquiry. Semantic (meaning) approaches to the study of language were disregarded, however, by the American structural linguists who so strongly influenced second language teaching in the mid-twentieth century. For structuralists, attempts to _interpret_ an utterance, to put it in a context with considerations of _who_, _when_, _why_, etc., lay outside the realm of theoretical linguistics proper. Thus it was that formal analysis--i.e., the analysis of the surface structure of language--would provide the basis for the teaching and testing materials that were developed in the 1950s and 1960s and are still in widespread use today. A concern for communicative competence, however, has brought us face-to-face with the contexts in which language is used. Once meaning is taken into account, matters of negotiation and interpretation are seen to be at the very heart of a communicative curriculum. Language in use, that is, language in context or setting, can no longer be ignored.

Central to any discussion of communication is the matter of mutual intelligibility. People learn languages to be able to communicate appropriately--that is, in a socially acceptable manner--with others. Sociolinguistics contexts of language use, social attitudes that determine preference for one or another linguistic norm, or language variety, and attempts to impose norms--to provide "linguistic watchdogs"--are among the topics that have been addressed by Kachru (1984). Kachru points out that the socially acceptable variety for a growing number of English language users in the world today is neither Received Pronunciation (RP), nor "BBC English," nor General American (GA), but one or a number of other regional norms. In such cases, adherence to external as opposed to local norms is attitudinally undesirable. Moreover, as the number of speakers of English as a second, third, or even fourth language continues to grow, the validity of the concept "native speaker" itself becomes doubtful. Matters of comprehensibility, acceptability, and tolerance of "error" take on important new dimensions in a world where nonnative varieties of English are used increasingly for communication across cultures and across languages. This example of the global spread of English provides a perspective on the nature of language use and language change in general, a perspective important to an understanding of the concept of communicative competence.

In my mind, the stated purpose of the TOEFL is to predict the nonnative speaker's ability to interpret, to express, and to negotiate meaning in U.S. college and university programs. This specified context implies an interaction primarily with native speakers of GA or with texts written or spoken for these same native speakers. However, a distinction should be drawn between the adoption of native speaker norms--writing or speaking like a native speaker of GA--and the achievement of mutual intelligibility--communicating with native speakers. A proper test of communicative ability in GA should reflect the latter rather than the former.

Efforts to represent the complexity of language use beyond the identification of discrete phonological, lexical, and syntactic units have been numerous. The componential model of communicative competence proposed by Canale and Swain (1980a) and Canale (1983), based on work by Halliday, Hymes, Munby, Savignon, van Ek, Widdowson, and others, represents one such by now well-known attempt. This particular model includes four components, summarized in Savignon (1983): (1) _grammatical competence_, (2) _sociolinguistic competence_, (3) _discourse competence_, and (4) _strategic competence_. Reference to these components has been useful in attempts to describe the more comprehensive nature of recent initiatives in language teaching and testing, e.g., Canale and Swain (1980b), Duran et al. (1985), Savignon (1983). However, as Oller (1978 and this conference) has repeatedly observed, a componential or taxonomic view of language use--whether the Canale and Swain model or lists of language functions, e.g., van Ek (1975)--cannot do justice to the integrative, compensatory nature of language proficiency. The whole is something other than the sum of its parts.

The division of language proficiency into dimensions of "basic," "communicative," and "autonomous" (Duran et al. 1984:14) seems similarly inadequate. In this representation, communicative language proficiency is described as a "form of social [dyadic] interaction in which the emphasis is normally less on grammatical forms and literal meaning than on participants and what they are trying to do." In contrast, autonomous language proficiency is described as "less directly social and more intrapersonal, representational (or self-directed) uses of language such as problem-solving, organizing one's thoughts, verbal play, poetry and personal writing." The suggestion here is that grammatical competence and discourse competence are more important in the latter than in the former and that therefore one cannot adequately test communicative competence through so-called autonomous tasks.

A more comprehensive perspective on the nature of communicative competence is that offered by Oller (this volume): "The critical elements of the dyadic definition are present within the experience of a single individual. In other words, a dyad of persons is not required. Rather, what is necessary in order for communication to occur is a dyad of semiotic systems plus an intelligent interpreter who is able to translate between the systems." Halliday (1978, cited in Berns, 1984) presumably had a similar view of language proficiency in mind when he observed that "Language comes to life only when functioning in some environment.... We do not experience language in isolation--if we did, we would not recognize it as language--but always in relation to a scenario, some background of persons and actions and events from which the things which are said derive meaning" (p. 28).

Reference to components is perhaps useful if our concern is with a descriptive model of communicative competence, a representation of the types of knowledge and skills needed for communication in a given setting. Such a <u>descriptive</u> model provides a basis for a <u>working</u> model that attempts to show how components of communicative competence are interrelated psychologically to form a set of statistically identifiable factors. (See Cziko, 1984.) However, it remains to be shown that the components used by theorists to describe language are psychologically distinct, suggesting that a learner can learn them separately or can learn more of one than another. Cziko (1984), Oller (1981), Bachman and Palmer (in press), and Bachman (this volume) provide a review of the major research that has been done along these lines to date and of the debate over the interpretation of findings.

Related to the problem of developing a model of communicative competence as a basis for evaluating English language proficiency is that of determining the extent to which deviations of various kinds from native speaker norms interfere with mutual intelligibility. Psycholinguistic studies provide ample evidence that utterances may be interpretable without being "natural," i.e., native-like, and that semantically deviant utterances are more likely to be misinterpreted than are grammatically deviant utterances. (See, for example, Khalil, 1984.) These findings, too, should be kept in mind in considering the validity of the TOEFL as a measure of a candidate's ability to interpret, express, and negotiate meaning in American academic settings.

II. THREE R'S OF THE TOEFL: RATIONALE, REALITIES, AND RESPONSIBILITY

> The Intensive English Institute (IEI) students are usually obsessed with the TOEFL and the necessary score that will get them admitted to an American university. They want no part of games, puzzles, role playing, etc. because these do not resemble *learning* to them. On one occasion I had a student refuse to finish a test constructed to test communicative competence because it didn't have face validity. If they are not filling in the oval-shaped marks on an answer sheet or completing a multiple-choice test, most students in the IEI feel they are being cheated and are not being adequately prepared to score a minimum of 500 on the TOEFL. (EFL Teacher, cited in Burke, 1984)

Savignon and Berns (1984) document several recent attempts to develop and to implement teaching and testing programs that represent a shift from the purely academic to the more practical aspects of language as communication. Among them are immersion programs in Ontario, curriculum revision in Finland, the RSA Oral Examination in Britain, and a foreign language proficiency requirement at the University of Pennsylvania. In each case, teaching and testing have gone hand in hand. It does little good to develop innovative curricula that place emphasis on communicative language use if evaluative procedures that reflect a similar emphasis are not in place. Conversely, testing programs that emphasize communication are valued as much for the impact they have on instructional programs as for the information they yield regarding learner achievement.

Swain (1984) describes the testing units that she and her colleagues in the Modern Language Centre of the Ontario Institute for Education have developed for use in province-wide assessments of the communicative performance of immersion students. A guiding principle in test construction was to work for "backwash," with backwash referring to the effect a test has on teaching practices:

> It has frequently been noted that teachers will teach to a test; that is, if they know the content of a test and/or the format of a test, they will teach their students accordingly. This is not particularly surprising, given the frequency with which educational administrators use tests, legitimately or not, to judge teacher effectiveness. (p. 196)

A similar point is made in the preface of the ARELS Oral Examination, a test of English language proficiency for foreign applicants to British universities:

> From the beginning the examination was regarded as a means as well as an end in itself. That is to say, it was considered no more important than the changes it would generate in language training by directing it towards modern and practical needs. A full-scale examination in spoken English would introduce new criteria, and influence beneficially classroom teaching and course planning. Its existence would encourage a shift from the purely academic to the more practical aspects of language as a means of communication. (Savignon, 1983:276)

It would thus seem of major importance in looking at the meaning of communicative competence in relation to the TOEFL to consider the impact of this widely-used test on English language programs around the world. Along with the rationale and the realities of the TOEFL, there is the responsibility of the TOEFL.

The 1983 edition of the TOEFL Test and Score Manual states unequivocally that TOEFL scores should not be used to predict academic performance. (See Hale, Stansfield, and Duran, 1984, for a review of research showing that the TOEFL is not a good predictor of academic performance.) The manual does claim that TOEFL is a measure of English proficiency and can assist an institution in making decisions as to an applicant's eligibility to begin an academic program. Data provided on institutional use of test scores show 525 to be a widely accepted minimum required for unrestricted acceptance into an academic program. The general trend in recent years is also reported to be toward raising rather than lowering this required minimum, although no reason for this trend is suggested. An increase in the number of foreign applicants is one possibility.

Given the widescale use of the TOEFL, one can easily understand the motivation of EFL students to perform well on this multiple-choice test and their general suspicion of classroom activities and evaluations that may not seem to them to be directly related to the tasks it presents. Thus, revision of the TOEFL to make it more reflective of the communicative competence of examinees would not only increase the validity of the test itself as a measure of English proficiency, but it would give a considerable boost to those EFL programs that seek to emphasize the practical use of English--the interpretation, conveyance, and negotiation of meaning, as opposed to the analysis and manipulation of formal features. All too often interest in new curricula and teaching materials that reflect a more communicative view of language is thwarted by what teachers and program administrators refer to as the "tyranny of the TOEFL." In the long run, the frustration felt by these teachers and administrators increases public criticism of the quality, meaning, and use not only of the TOEFL, but of Educational Testing Service tests in general, e.g., Owen (1983):

> [The ETS budget] comes from people required to pay for the privilege of submitting to ETS exams in order to pass various checkpoints in America's social hierarchy. ...ETS floods institutions with statistics in order to make itself seem indispensable and to uphold the "scientific" facade it has erected around its tests. If the colleges had to pay, few of them would bother. (p. 37)

It is therefore in the best interests of all to make criteria of evaluation reflective of the functional proficiency that an increasing number of instructional programs, both academic and nonacademic, are setting as their goal.

The obvious first step in revising the TOEFL is to provide not only more context, but more cohesive and coherent context--i.e., a world of experience, and to specify factual domains appropriate to examinees. (See Duran et al., 1985, Oller, this volume; and Douglas, this volume.) The findings of the review of TOEFL items reported in Duran et al. (1985) are important in this respect:

> ...the range and complexity of skills required on various sections of the TOEFL was directly related to the amount of language and to the semantic and textual complexity of TOEFL items. The more language used, and the more authentic the language, the greater the number and kinds of communicative skills required of examinees. (p. 44)

The following sample items from Section I of the TOEFL, Listening Comprehension, the section given greater emphasis in admissions decisions by 41 percent of the institutions surveyed by the TOEFL program in the fall of 1982 (TOEFL Test and Score Manual, 1983), illustrate the prevailing emphasis on form rather than function. All examples are from disclosed test form 3ELTF12, used on December 10, 1982.

In Part A, examinees hear a single utterance, spoken once, and are asked to decide which of the four sentences in their test booklet is closest in meaning to the statement they have heard.

1. She sat down and made herself comfortable.

    (A) She made it while she was sitting down.
    (B) She built the chair and table herself.
    (C) She settled into a comfortable place.
    (D) She liked to sit by herself.

In this first item, the absence of all context and the lengthy responses from which to choose forces examinees to concentrate unnaturally on the form as opposed to the meaning of the spoken statement. While the statement suggests that a girl or woman is settling into a comfortable chair, perhaps in a home or a waiting room, this image is blurred by the first two distractors, which do indeed distract even an alert native speaker. Nor is the image of a girl or woman sitting in a chair immediately re-evoked by (C) since the use of the word place suggests a residence--e.g., She's got a nice place. While the correct response can eventually be selected through the process of elimination, this requires considerably more than the interpretation of the spoken statement.

4. The children walked into the museum two by two.

(A) Two children entered the museum.
(B) There were two children in the museum.
(C) The children entered the museum in pairs.
(D) The children should go into those two museums.

The spoken statement sounds unnatural to begin with. Use of the expression two by two to refer to animals entering Noah's ark would be more appropriate, perhaps, but not very relevant to examinees' needs. In my experience as a teacher's helper on class excursions, presumably the setting evoked here, children may be asked to form a double line, with the result that they would pass through the door of a museum two at a time. The repeated use of the words museum, two, and children in the options makes for quite a scramble of images, again forcing examinees to focus on form. In sum, the item looks as though it was contrived to see if an examinee could equate two by two with in pairs, without careful consideration of the authenticity of the surrounding language. Although the lexical items were "contextualized," the basic nature of the task remains discrete point.

The same kind of formal focus is apparent in the immediately following item where the word pair again appears, this time with another meaning, along with repair for added confusion. Selection of the correct response hinges on equating repair with fix.

5. Bob had the shoemaker repair his sandals.

(A) Bob bought a new pair of sandals.
(B) Bob's sandals were fixed.
(C) The shoemaker only made sandals.
(D) The shoemaker wore sandals.

A final example of the emphasis on form rather than substance in Part A of Listening Comprehension is item 8.

8. Joe would rather wear the blue coat than the grey one.

(A) Joe doesn't like the grey coat as much as the blue one.
(B) Joe asked where the blue and grey coats were.
(C) Joe would like to wear the blue and grey coat.
(D) Joe didn't want the blue coat or the grey one.

Since no context is provided, it's not at all clear why Joe wold rather wear the blue coat, and the repetition of the words blue and grey in each response is bound to increase an examinee's confusion and anxiety. The image conjured up for me is that of the American Civil War. (Perhaps the item writer had used a text on that topic as a basis for items in another section!) Among the questions I would like answered are: Is Joe a boy or a man? Is the coat a sportcoat or an overcoat? Why does Joe prefer the blue coat on this occasion, since both coats are presumably his? Here, as in Part A generally, more sustained texts with the kind of redundancy found in natural speech would seem a better basis for evaluating examinees' ability to interpret spoken General American English. (See Savignon, 1983, pp. 249-254, for further illustration of the distinction between form and substance in test item analysis.)

Part B, consisting of short conversational exchanges, does provide more context. Selection of the intended response sometimes hinges on the comprehension of but one utterance (see Duran et al. 1984), as it does in Part A. In general, however, these items have sufficient context and redundancy to allow examinees to interpret meaning. The generally shorter and more varied responses further enhance the quality of items in this part.

24. It's really steamy today. The temperature must be over ninety.
Yes, I know. There's a lot of moisture in the air.
What does the woman say?

(A) Steamed rice would be nice for dinner.
(B) There are a lot of hot-air balloons.
(C) Steve must be overy [sic] ninety years old.
(D) It's hot and humid outside.

23. These look like good grapes.

They taste sour now, but the wine we make from them should be just fine.

What are these people discussing?

(A) The injustice of the fine.
(B) Their wine grapes.
(C) Their whining son.
(D) The coming spring shower.

The formal delivery and the rather contrived nature of some of the "conversations"--read as if they were "test" items--does reduce the authenticity of the language, e.g., item 25:

25. There's Bill on his motorcycle. Did he take it to the garage to be fixed?

Don't be silly; that would have been a waste of money. It only had a flat tire.

But this seems a simple matter to correct. In my view, more items of this kind and the elimination of Part A would improve the overall validity of this section.

In general, I concur with the observations made by Duran et al. (1985) and Oller (this volume) regarding Part C of this section, minitalks. These longer samples of spoken English have considerable potential for involving an examinee's communicative competence. The use of a formal expository style similar to that encountered in American academic settings seems appropriate, moreover, given the intended purpose of the TOEFL. However, in vocabulary, sentence formation, and discourse structure, many of the texts presented more closely resemble written expository prose than oral classroom exposition. Some actual samples of classroom exposition, including the pauses and repetitions that commonly occur, would no doubt help to give test item writers a better sense of the nature of such discourse. In addition, an effort should be made to provide texts that are both interesting and varied. Academic need not mean dull.

In my view, Section II, Structure and Written Expression, is the least defensible of all sections. Keeping in mind that communicative competence requires the achievement of mutual intelligibility, not the adoption of native-speaker norms, and that such mutual intelligibility is less affected by grammatical deviance than by semantic deviance, it is particularly difficult to defend the content and format of the discrete-point structural items in this section. They serve only to reinforce a classroom emphasis on form, as opposed to function.

I further agree with Oller that the test instructions themselves could be greatly simplified. They should be straightforward, nonrepetitive, and readily understood, with help from a few simple examples, by examiness with moderate English language proficiency.

To be sure, there are a number of practical considerations, including budgetary constraints and the requirements of large-scale administration, that must be taken into account in recommending changes in content and format. Keeping these constraints in mind, those of us participating in this conference have been challenged to discover ways in which the TOEFL, one of the most highly respected of language tests, and thus one that influences English language teaching around the world, can be revised to present tasks that are both natural and interesting to examinees, that engage their ability to interpret and convey meaning. In sum, they should be tasks that encourage examinees to make use of their communicative competence.

References

Bachman, L. F., & Palmer, A. S. (in press). Basic concerns in language testing research. Reading, MA: Addison-Wesley.

Berns, M. (1984). Functional approaches to English: Applications and implications. Paper presented at the AILA Meeting, Brussels.

Breen, M., & Candlin, C. N. (1980). The essentials of a communicative curriculum in language teaching. Applied Linguistics, 1, 89-112.

Burke, K. (1983). Personal communication.

Canale, M. (1983). From communicative competence to communicative language pedology. In J. Richards and R. Schmidt (Eds.), Language and Communication (pp. 2-27). London: Longman.

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

Canale, M. & Swain, M. (1980b). A domain description for core FSL communication skills. The Ontario assessment instrument pool: French as a second language, junior and intermediate divisions (pp. 27-39). Toronto: Ontario Ministry of Education.

Cazden, C. (1976). How knowledge about language helps the classroom teacher--or does it: a personal account. Urban Review, 9, 74-90.

Cziko, G. (1984). Some problems with empirically-based models of communicative competence. Applied Linguistics, 5, 23-38.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. (1985). The TOEFL from a communicative viewpoint on language proficiency: A working paper (TOEFL Research Report 17). Princeton, NJ: Educational Testing Service.

Goleman, D. (1984, July 31). Style of thinking, not I.Q., tied to success. New York Times, p. 15.

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). Summaries of studies involving the Test of English as a Foreign Language, 1963-1982 (TOEFL Research Report 16). Princeton, NJ: Educational Testing Service.

Halliday, M. (1978). Language as a social semiotic. Baltimore, MD: University Park Press.

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

Kachru, B. (1984). Regional norms for English. In S. Savignon & M. Berns (Eds.), Initiatives in communicative language teaching (pp. 55-78). Reading, MA: Addison-Wesley.

Kelly, L. (1969). 25 centuries of language teaching. Rowley, MA: Newbury House.

Khalil, A. (1984). Communicative error evaluation: A study of American native speakers' evaluations and interpretations of deviant utterances written by Arab EFL learners. Unpublished doctoral dissertation, University of Illinois: Urbana-Champaign.

Lerner, B. (1979). The war on testing: Detroit Edison in perspective. Princeton, NJ: Educational Testing Service.

Lyons, R. (1984, June 19). AMA orders inquiry into test for license. New York Times, p. 21.

McClelland, D. (1973). Testing for competence rather than for 'intelligence.' American Psychologist, 28, 1-14.

Oller, J. (1978). Pragmatics and language testing. In B. Spolsky (Ed.), Approaches to language testing (pp. 39-57). Arlington, VA: Center for Applied Linguistics.

Oller, J. (1980). Language testing research. In R. Kaplan (Ed.), Annual review of applied linguistics - 1980 (pp. 124-150). Rowley, MA: Newbury House.

Owen, D. (1983, May). The last days of ETS. Harper's, pp. 21-37.

Savignon, S. (1972). Communicative competence: An experiment in foreign language teaching. Philadelphia: Center for Curriculum Development.

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

Savignon, S. & Berns, M. (Eds.). (1984). Initiatives in communicative language teaching. Reading, MA: Addison-Wesley.

Saville-Troike, M. (1984). Ethnography and linguistic research. Paper presented at English as a second language colloquium. Urbana-Champaign: University of Illinois.

Swain, M. (1984). Large-scale communicative language testing: A case study. In S. Savignon & M. Berns (Eds.), Initiatives in communicative language teaching (pp. 185-201) Reading, MA: Addison-Wesley.

van Ek, J. (Ed.), (1975). Systems development in adult language learning: the threshold level in a European unit credit system for modern language learning by adults. Strasbourg: Council of Europe.

A RESPONSE TO SANDRA SAVIGNON'S "THE MEANING OF COMMUNICATIVE COMPETENCE IN RELATION TO THE TOEFL PROGRAM"[1]

Diane Larsen-Freeman

The TOEFL Program is to be commended for convening this gathering. As others here have noted, the TOEFL has been a standard setter since its inception and, with the responsible attitude manifest here, it will doubtless continue in its leadership role.

My charge was to react to Sandra Savignon's contribution to this conference. I will do this by focusing my remarks on a theme that is recurring in her paper and the one to which she devotes the most space.

Savignon writes:

> In my mind, the stated purpose of the TOEFL is to predict the nonnative speaker's ability to interpret, to express, and to negotiate meaning in U.S. college and university programs. This specified context implies an interaction primarily with native speakers of GA or with texts written or spoken for these same native speakers. However, a distinction should be drawn between the adoption of native speaker norms--writing or speaking like a native speaker of GA--and the achievement of mutual intelligibility--communicating with native speakers. A proper test of communicative ability in GA should reflect the latter rather than the former. (p. 21)

And then again:

> Related to the problem of developing a model of communicative competence as a basis for evaluating English language proficiency is that of determining the extent to which deviations of various kinds from native speaker norms interfere with mutual intelligibility. Psycholinguistic studies provide ample evidence that utterances may be interpretable without being "natural," i.e., native-like, and that semantically deviant utterances are more likely to be misinterpreted than are grammatically deviant utterances. (p. 22)

[1] I would like to thank my colleague Donald Freeman for his helpful comments on this paper.

And finally:

> Keeping in mind that communicative competence requires the achievement of mutual intelligibility, not the adoption of native-speaker norms, and that such mutual intelligibility is less affected by grammatical deviance than by semantic deviance, it is particularly difficult to defend the content and format of the discrete-point structural items in this section. They serve only to reinforce a classroom emphasis on form, as opposed to function. (p. 28)

I take exception to such statements. I realize that challenging the dominance of structure/form is a popular theme these days, and I find the role of reactionary uncomfortable; nevertheless, I truly believe that removing grammatical accuracy from an assessment of communicative competence would be a grave mistake. I think most language professionals agree that all languages consist of at least three interacting dimensions: form, meaning, and pragmatics. Any approach to language teaching or testing that elevates one or two of these dimensions and systematically ignores or subordinates the remaining dimension(s) will prove inadequate. Furthermore, the approach will be subject to the caprice of fashion, just as such approaches have been for many years. To my mind, the value of communicative competence has been that it forces us to broaden our view of language use in context; it has not replaced the assumption that the accuracy of linguistic form is an essential component of language use.

Second, even if we were to agree to ignore the form of the test candidates' productions and assess instead their ability to achieve mutual intelligibility, I fail to see how any standards could be established. Intelligibility surely is a relative phenomenon dependent upon who the interlocutors are, what their relationship is, how much information is shared, etc.

Furthermore, while I am sympathetic to the notion of there being many "Englishes," the TOEFL is, after all, a measure of the ability of the examinees to use the English associated with a particular setting--an American academic context. It therefore makes sense to me that the TOEFL assesses the candidates' ability to use the English appropriate to this context.

Since my custom is to play the doubting game before the believing game (cf. Larsen-Freeman, 1983), what I have asserted so far was my initial reaction to Savignon's comments. When I then switched perspectives and tried to detach myself from my initial reaction, I realized that I could agree with the spirit of her remarks, if not the form. It has been widely observed that a person can express his or her ideas very forcefully and yet exhibit linguistic deficiency (in the sense of failing to conform to native-speaker norms) at the same time. It is perhaps less widely acknowledged, but from my perspective equally valid, that a person might also be able to function perfectly well in a community where the target language is used but might be deficient in conforming to certain sociolinguistic norms or discourse constraints. That is, a level of English necessary to function in an American academic context may still be achieved even without a person's fully controlling all aspects of the target language.

Assessing this functional ability, in my opinion, is the challenge that research in the area of communicative competence issues to the TOEFL program.

Let me put it another way: We know the TOEFL works well psychometrically. Candidates who receive a high score on the examination by and large do have the English ability to equip them for study in an American academic setting. We don't necessarily know the reason for its success, however.

It may work because, by design or by good fortune, the TOEFL is already a good test of communicative competence. I note that Oller (this volume) ends his paper by observing that "...everything points to the conclusion that the TOEFL is presently a fairly good measure of communicative competence..." (p. 149).

Then again, the TOEFL may work not because it is a particularly good measure of communicative competence as we understand it today, but because candidates who are linguistically competent enough to receive a high score on the test, are also sufficiently communicatively competent to study in this country. I note that Douglas (this volume), in discussing the TSE, observes that "what communicative aspects there may be in the tasks are largely lost in the scoring procedure, which requires scorers to ignore content and focus on primarily the linguistic features of the responses" (p. 172). Still, despite this, Douglas later observes (p. 172), "I can testify that the scoring procedure is quite workable and does seem to produce an evaluation that bears some relationship to what one intuitively feels to be a candidate's strengths and weaknesses in oral production." Thus, there may be an asymptotic relationship between linguistic competence and communicative competence.

Whatever the reason for why the TOEFL works, the point is it works... or does it? It works in the sense that someone who does well on it may be communicatively competent, but does it work the other way? Can we say that someone who does not receive a high TOEFL score is not qualified to pursue study in an academic setting in this country?

I am not sure that a theory of communicative competence will ever answer that for us. I despair of ever reaching accord on a definition of communicative competence (see Rivera's comment cited in Oller [this volume]). I also doubt we will ever complete a list of factors comprising it. Canale and Swain (1980) in their original descriptive model suggested there were three. Bachman and Palmer (1982) in their working model (following Cziko's [1984] distinction between descriptive and working models) found three also: one general and two specific trait factors.

Canale (1983) reports four. Larsen-Freeman (1981), in reviewing the research literature, unearthed five areas that appeared to her to make up communicative competence.

Although identifying components of communicative competence is a worthwhile endeavor, the task may be boundless (Richard Duran, personal communication). What are the parameters of one's communicative competence?[2] Moreover, identifying more and more factors that comprise communicative competence or regrouping them into different configurations does not mean that we have increased our understanding. A pie is still a pie, no matter the size or the shape of its slices.

Nevertheless, despite our theoretical shortcomings of the moment, what I think we would agree on is that no one would want to exclude anyone from the label "communicatively competent" who is an effective communicator but who has an incomplete knowledge of discourse features, sociolinguistic norms, linguistic forms, or whatever. In order to avoid excluding such a person, the only recourse it seems to me is to have a section of the TOEFL devoted to candidates' demonstrating what they do know, not being penalized for what they don't know (Gattegno, 1976).

How would this be accomplished? I don't think it can be accomplished by improving the contexts for items, lengthening them to include more information in order to tap discourse processing, or attempting to make the language and its delivery more authentic. With all due respect to my colleagues at this conference who have made these recommendations, these are all worthwhile and will presumably enhance the exam's effectiveness as a measure of communicative competence. For what I have in mind, however, these are insufficient. In order to "bias for the best" (Duran et al., 1985), I think we need to let the candidate show us the best that he or she can do.[3] What we need, I think, is some sort of performance (I presume written) task.

As to what exactly the task would be, I would like to see research of the sort that Angelis (1982) and Bridgeman and Carlson (1983) conducted.[4] Let us find out precisely what sort of discourse domains (Douglas, this volume) and tasks our TOEFL candidates will encounter during their academic careers, and let us design tests to reflect these. I realize that the most difficult (impractical) aspect of what I am proposing is in the evaluation (scoring) of the candidates' performance. I think we can rely on research to help us with this challenge also.

[2] See, for example, John Oller's paper in this volume. His definition is much broader than other ones previously considered.

[3] There is a precedent for this at ETS, by the way, in the Advanced Placement exams. (I thank Donald Freeman for bringing this to my attention.)

[4] At the session in Princeton, I learned that Don Powers is also carrying out research of this sort in the area of listening comprehension.

Bridgeman and Carlson (1983) asked faculty members how they evaluated their students' writing. They reported that "they relied more on discourse level characteristics than on word or sentence level characteristics" (p. iii). I think we need to learn more about what they, and others like them, do. If there is no consensus, it may come down to leaving it up to the respective institutions to evaluate a candidate's performance according to their own criteria, with ETS acting as a broker to administer the performance task and to see that the appropriate institution receives the candidate's actual written product.

I recognize that no matter what we come up with, as Bachman (this volume) reminds us, the performance task will still be an indirect measure of the competence in which we are interested; still, having a task such as the one I envision moves us in the direction of the two principles (with which I am delighted) stated at the end of the Duran et al. (1985) study:

> Bias for the best
> Bias for rewarding feedback.

It may be that all I am calling for is that we obtain some measure of strategic competence, an important component of communicative competence, which Duran et al. (1985) did not evaluate since they felt it was not being tapped by the present form of the TOEFL. It may, however, be being assessed after all, if one adopts Larsen-Freeman's (1981) definition and allows it to apply to receptive skills as well:

> Strategic competence is a dynamic process. It is a superordinate process responsible for controlling the smooth flow of communication. It enables the participant in discourse to draw upon his or her knowledge in the other four areas and to put this knowledge together in a fluent, creative way--as a listener, speaker, reader or writer. (p. 118)

I do note, however, that when Bachman (this volume) did evaluate the TOEFL in light of whether or not its tasks involved communicative language performance, there was only one item type in which he reported that strategic competence was required. This was in the reading comprehension section.

It may be, therefore, that if the TOEFL is remiss at all, it is remiss in failing to tap this basic competence. On the other hand, a performance task may be a measure of more than strategic competence. It may assess other aspects of communicative competence--the quality of ideas, for example--as well.

In any event, I think there are added benefits to be derived in moving more in the direction of employing "direct"-performance-based tests.[5] For one thing, such a move would be a direct contribution to a theory of communicative competence. Rather than the TOEFL program fearing that its test fails to keep abreast of the researchers' understanding of communicative competence, it could take the lead in helping us increase our understanding.

Second, it could move the test in a direction that the Duran et al. (1985) study recommended: that of a criterion-referenced test. Such a move could have a most welcome positive backwash effect, I believe.

Third, it might make the "meaning" of a TOEFL score more obvious. Currently the program receives queries from admissions officers who want to know what a certain score on the TOEFL means a candidate can do. A criterion-referenced performance task would help in this regard.

Finally, moving in this direction defines a research agenda that I originally proposed at a TOEFL Research Committee meeting in 1982. At that meeting the committee was presented with a grid that stated the goals of TOEFL research on the horizontal axis and types of validity studies on the vertical axis. On the horizontal axis were: "to maintain and improve the quality of the TOEFL, to support the effective and appropriate use of the TOEFL, to develop and refine measures of English language skills suitable for foreign students, to assist academic institutions in examining the process and outcome of education for foreign students." What I proposed at that meeting (and which may have been included since) was that a further objective should be added to the list on the horizontal axis--that of improving the TOEFL for the test taker.

In sum, what I have been arguing as a response to Savignon's paper is that we not discontinue assessing a candidate's ability to control linguistic form. This component of communicative competence should rightfully be assessed, in my opinion. What we must avoid, however, is limiting the parameters of our evaluation and thereby discriminating unfairly against someone who is an effective communicator, despite certain deficiencies in his or her ability to control all aspects of the target language.

Just as the effect of the notion of communicative competence has been to broaden our view of language in use, so I am calling for a broadening of the parameters of the test. By placing emphasis on the correctness of a candidate's linguistic performance, does the TOEFL now unduly limit our view of a candidate's proficiency? It seems to me that this is the challenge of communicative competence that should command the attention of the TOEFL program.

[5] Even though Gordon Hale tells us that there is a relationship between direct tests and the TOEFL.

References

Angelis, P. (1982). Academic needs and priorities for testing. American Language Journal, 1, 41-56.

Bachman, L. F., & Palmer, A. S. (1982). The construct validity of some components of communicative proficiency. TESOL Quarterly, 16(4), 449-465.

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

Canale, M. (1983). Program evaluation: Where do we go from here? Plenary address presented at the TESOL summer meeting, Toronto.

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

Cziko, G. (1984). Some problems with empirically-based models of communicative competence. Applied Linguistics, 5(1), 23-38.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. (1985). The TOEFL from a communicative viewpoint on language proficiency: A working paper. (TOEFL Research Report 17). Princeton, NJ: Educational Testing Service.

Gattegno, C. (1976). The common sense of teaching foreign languages. New York: Educational Solutions, Inc.

Larsen-Freeman, D. (1981). The WHAT of second language acquisition. Plenary address at the Detroit TESOL Convention. In M. Hines & W. Rutherford (Eds.), On TESOL '81 (pp. 107-128). Washington, DC: TESOL.

Larsen-Freeman, D. (1983). Second language acquisition: Getting the whole picture. In K. Bailey, M. Long, & S. Peck (Eds.), Issues in second language acquisition research (pp. 3-22). Rowley, MA: Newbury House.

EXPLAINING COMMUNICATIVE COMPETENCE
LIMITS OF TESTABILITY?

Christopher N. Candlin

PREAMBLE

I wish to offer an analysis of communicative competence that is, in two key respects, an alternate perspective to the one that seems currently to be held in the applied linguistics literature. The first of these two respects concerns the relationship between knowledge and ability; the second has to do with the capacity to explain as well as to describe and interpret utterances. I will argue that while a traditional analysis of communicative competence requires those involved in testing merely to make changes in the range and scope of their view of language, what I propose involves radical changes in the kind of view they take. Finally, what I propose may be important to the approach we adopt toward the evaluation of communicative competence, particularly within a TOEFL framework.

INTRODUCTORY

It is not at all surprising that in the last five years there has been an upsurge of concern in language testing circles about the concept of communicative competence, as key publications such as Canale and Swain (1980), Alderson and Hughes (1981), Canale (1983), Rea (1985), Swain (1985), and Douglas and Selinker (1985) amply testify. There are, I believe, three general and three testing-specific reasons for this concern.

Among the general reasons, first, there is the recognition of a historical dimension, the requirement that innovations in language testing ought not to be at odds with developments in linguistic theory and description that they subsume. Second, there is a disciplinary dimension that sets a boundary around what is regarded as testable, in terms of testing's view of what is contained within "language study"--however broadly or narrowly defined. Finally, there is what we call in critical applied linguistics terms a reflexive dimension; we argue that the connections between language and social life cannot be ignored. Our view of language (and hence our view of language testing) must maintain a connection between the social identity and purposes of its users and the social formations reflected and maintained in their modes of language use. Insofar as communicative competence looms large in our consideration of these three general dimensions, it is difficult for language testing to be unconcerned. To language testing specifically, there are immediate consequences for our interpretation of the concept of validity.

As to the three testing-specific reasons, in terms of content validity, tests need to measure communication. In terms of construct validity, they should test "communicatively." Their concurrent validity will be measured against their adequacy in the posttesting work- or study-place where "communicatibility" is at a social and personal premium.

I have, however, been communicating rather loosely. It is time for some definition.

1. What do we mean by communicative competence?

Let me begin with two quotations:

> From the speaker's point of view the problem is: 'Given that I want to change the state of the hearer in such a way, how do I frame my utterance in order to make that outcome most likely?' From the hearer's point of view, it is 'Given that the speaker said so-and-so, what is the most likely reason for his saying so-and-so?' (Leech, 1979, p. 420)
>
> ...(communicative competence is) the knowledge of discourse processing conventions and related communicative norms that participants must control as a precondition to being able to enlist and sustain conversational cooperation. (Gumperz, 1984, p. 280)

If we unpack these quotations, we become aware of the analytical and applicational problems in answering this question. I shall identify two.

(1) Both quotations see communication as an interactive and cooperative process where understanding emerges and is jointly managed, rather than inherently available in the semantics of the individual words employed.

(2) Both quotations imply an uncertainty of perspective; do we see the speaker's and hearer's target as one of conforming to external controls or rather of creatively exploiting convention to create meanings, and, thereby, new conventions?

We might also add that both suggest that this communicative activity cannot be exclusively social or psychological, but rather socio-psychological, where the integration of knowledge and ability is not at all clear in any given encounter.

In essence, my argument is that a traditional (descriptive) view of communicative competence drives us toward conformity to norms (Oller's "fixing the facts" [Oller, this volume]), whereas an alternative (explanatory) view encourages us toward a critical creativity. My sympathies lie with this second view, understanding communicative competence as capacity (in Widdowson's sense [Widdowson, 1983]): the ability to create meanings by exploring the potential inherent in any language for continual modification in response to change, negotiating the value of convention rather than conforming to established principle. In sum, I favor a coming together of organized knowledge structures with a set of procedures for adapting this knowledge to solve new problems of communication that do not have ready-made and tailored solutions.

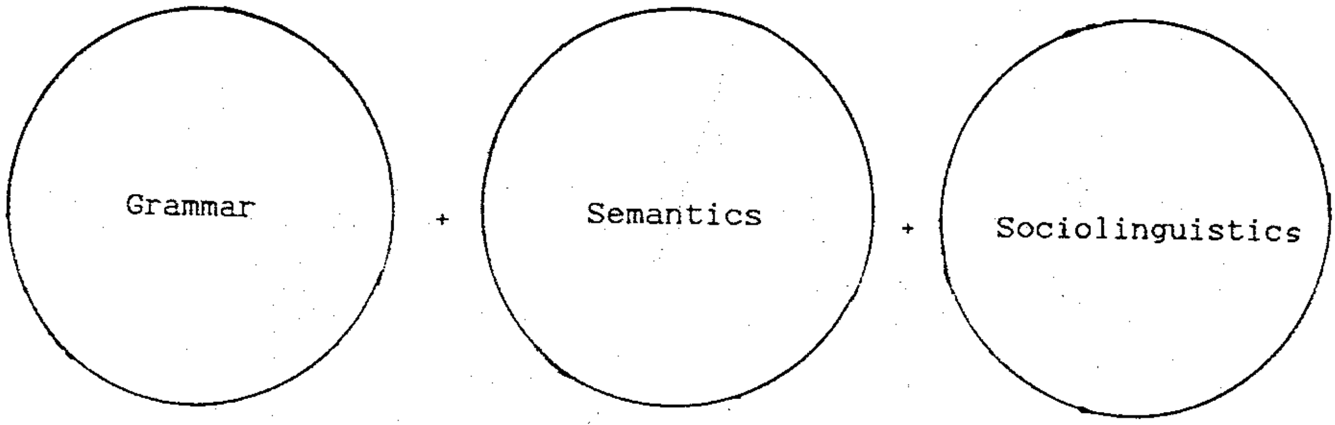
I shall, however, argue that this orientation toward creativity and nonconformity, towards diskurs rather than handeln in Habermas' sense (Habermas, 1970), throws a burden upon our capacity to describe and interpret the products of an individual's communicative competence and, also, to explain it. Such a burden falls also upon the applications of any such description/interpretation and explanation, hence the need for my subtitle: limits of testability?

We ought, perhaps, first to step back from the brink of this serious abyss of empirical problems and take a historical, a disciplinary, and a reflexive look over the edge.

In 1966, during a debate at the Yale Linguistic Club on the boundaries to be placed on linguistic theory, Dell Hymes saw his task as one of augmenting the scope of competence by discussing how ways of speaking depend on the culture of speech events--both their structural and stylistic codes and their appropriate modes of behavior. As such, communicative competence requires not only an understanding of grammar and semantics, but also of sociolinguistics, particularly in terms of the structures and occurrence of social events and how utterances realize social roles and achieve both social and communicative goals. This augmented potential is well documented in Hymes (1967, 1972) and in the following figure:

Figure 1

Hymes Augmented View of Communicative Competence



In essence this is a broadening of the scope of competence, constrained by the probabilities of occurrence of linguistic items and the capacity of the individual to control and process them.

An elaboration of this basic paradigm is provided in the paper by Canale and Swain (1980) to which much "communicative" testing owes its theoretical justification. It is my contention that such elaborations, while admittedly offering a more delicate enthnographic map for the classification of utterances, do not enhance our understanding (and our evaluation) of that communicative capacity to which I have referred. They do not because they offer no explanation of which aspects are used in particular circumstances and how they are drawn upon in the process of making meanings. It is worthwhile, in fact, considering the more recent disclaimers by both Canale and Swain on this matter:

"... the theoretical framework is not a model of communicative competence where model implies some specification of the manner and order in which the components interact and in which competencies are normally acquired"

"... the question of how these competencies interact with one another (or with other factors involved in actual communication) has been largely ignored here."

"Ultimately, it is a model of communicative competence that must be articulated for second language pedagogy (we may add 'testing' (CNC)), since a model has more direct applications than has a framework."

"...the view expressed here is a modular or compartmentalized one... communicative competence is analyzed as composed of several separate factors that interact."

"There is still relatively little empirical evidence for distinguishing the four areas of competence (i.e., grammatical, sociolinguistic, discourse and strategic competence [CNC]), proposed here."

"... the primary goal of a communicative approach must be to facilitate the integration of these types of competence for the learner" (Canale, 1983, pp. 12-13).

Furthermore, the five guiding principles enunciated by Canale in his 1983 paper, though focussed on language pedagogy, might equally be held to be criteria for language testing in a communicative framework, viz that tests should demonstrate coverage of the competence areas, satisfy the communication needs of the learner/examinee, provide meaningful and realistic interaction, draw consciously upon the learner's native language skills, and take advantage of a curriculum-wide approach. This would indeed be a charter for communicative language testing.

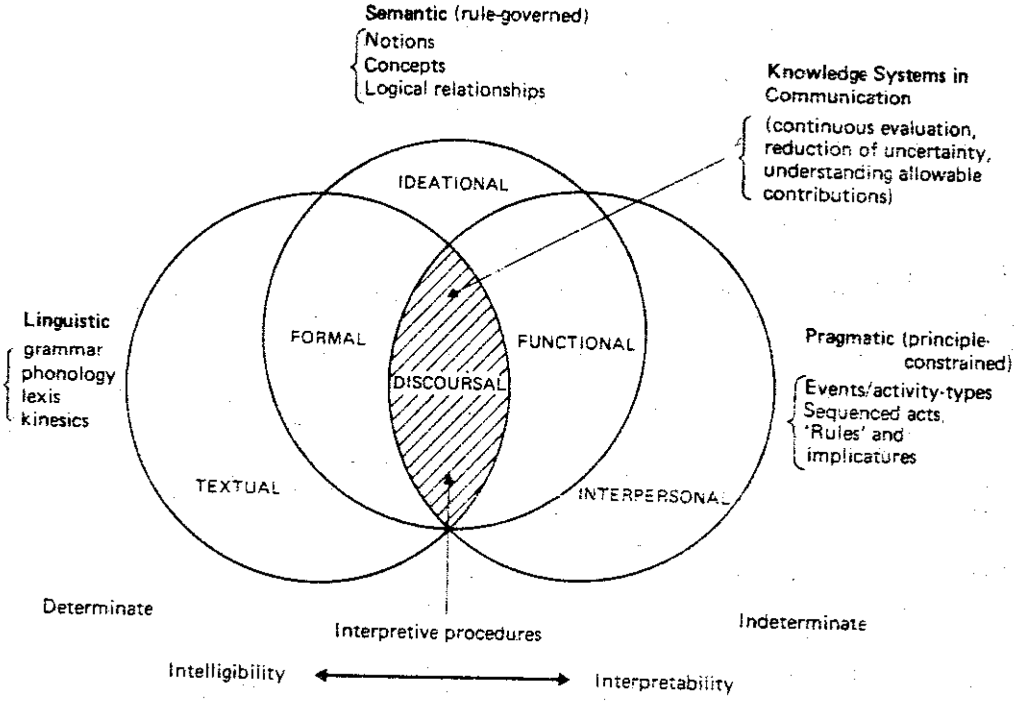
Swain, too, is cautious in her 1985 paper about our capacity to use the constructs of the 1980 model for assessing proficiency. For one thing, the validation of the model presupposes a capacity to adequately characterize the interaction between the components, a consideration that such a discrete "framework" seems to preclude from the outset. It would seem from the caveats of Canale and Swain that an alternative model is required, one that is more in harmony with the quotations with which we began and that addresses those desiderata cited above.

Just such an alternative is, I believe, offered in Figure 2, itself derivative of a Hallidayan view of language with its focus on the interdependence of text, ideation, and interpersonality. (Halliday, 1979)

Figure 2

A Hallidayan Model of Communicative Competence

From Language as Social Semiotic by M. A. K. Halliday, 1979, London: Arnold.



Problems facing Interpretability

1. Textual (Linguistic)
2. Ideational (Semantic)
3. Interpersonal (Pragmatic)

4. Discoursal "capacity" (Psycholinguistic)

The model has three presupposed characteristics:

(1) It explicity recognizes the interplay among the three worlds, lying behind any instance of text.

(2) It declares that the worlds of text and ideation--of sentences and sentence meaning--enjoy a determinism not appropriate to the negotiable world of interpersonality, i.e., that rules and principles are not identical.

(3) It is aware that the interpretation of utterances in discourse involves the speaker's and hearer's sociopsycholinguistic capacity to make meanings and thereby create convention.

This Hallidayan model is intended to make explicit that communicative competence is a matter both of knowledge (the three worlds) and ability--our capacity to interpret these worlds and, in so doing, to augment our knowledge through our expression. It adopts a negotiative rather than a conformist stance. It is, as Widdowson (1983) points out, ethnomethodological rather than ethnographic. In this alternative view, then, it is not only crucial to capture the systems of semantics and syntax (the schemata and the frames of interpersonality), but also to incorporate the variability inherent in human performative and interpretive capacity. Setting the procedures associated with ability to work on knowledge requires us at least to review our definition of each and their interrelationship in interaction. Knowledge structures are to be seen less as lists of items and more as organized and coherent experiences of texts, constellations of items and scripted routines; procedures are those on-line processing strategies of inference and judgment that make use of pragmatic principle to achieve interpretation.

We ought, before proceeding, to emphasize again the analytical and applicational difficulties, drawing on the arguments of Gumperz (1984) in particular:

(1) Although discourse is goverened by organizational principles, these do not operate like the categorical rules of grammar. They are more guidelines, maxims, and standards.

(2) As such, discourse cannot rely on generally shared conventions, but on those that are "differentially distributed in accordance with social (and we may add, 'ideological') boundaries" (Gumperz, 1984).

(3) Moreover, the textual realizations of these conventions are "transitory and difficult to acknowledge or retrieve" (Gumperz, 1984) and, accordingly, difficult to transfer and convey to others.

(4) If the context of interpretation is jointly and locally constructed, then it becomes difficult in any Labovian-correlational way to assign social values to decontextualized utterances.

These represent more than analytical and descriptive difficulties; they present substantial problems for any applications where the conditions of negotiation of value within a shared contextual world are not normally present, as, for example, in the world of language testing. It would seem, then, that we will have to offer some expansion to Alderson's classic (but underasked) question: "What is the test's view of language?" (Alderson 1981)

In particular, we shall need to accommodate (and be able to evaluate) what the learner presupposes and what the learner is capable of negotiating. Consider, for a moment, what this implies: our target becomes not only textual, ideational, and interpersonal knowledge, but also the learner's capacity for making meanings, drawing inferences, and negotiating a range of variable intentions, schemata, and routines. Indeed, looking ahead to examinee appraisal, we might want to regard highly someone who can make *any* meaning out of a culturally- and socially-specific text, and regard even more highly someone who could offer and evaluate a range of alternative readings. There would be correlates in performance, too: examinees who could compensate for especially wide gaps between their schematic knowledge and that of their interlocutors ought to be similarly rewarded. After all, they are in the commonsensical world of everyday communication. The following instances will perhaps bring home the issue, instances where our present lack of hard information exacerbates the difficulty:

(1) To what extent are the principles for utterance interpretation socially and culturally specific, and to what extent are they general?

(2) To what extent are there universal activity types and prototypical transactions?

(3) Do particular activity-types demand specific modes of interpretation?

(4) If such activity-types enjoy some universality, are they similarly structured discoursally?

(5) What social functions do universal acts (say, *questioning*) play in different languages?

(6) Do certain textual features hinder the activation of appropriate interpretive procedures for certain non-native speakers?

It is here that the *disciplinary* dimension becomes relevant, particularly in the need to incorporate within our definition of communicative competence not just the *linguistic* (in the broad Hallidayan sense), not just the *psycholinguistic* (implied in the interaction between knowledge and ability), but also the *sociological* and the *sociopsychological*. This is not only a matter of conviction, as Habermas indicates:

> ... a sociology that accepts meaning as a basic concept cannot abstract the social system from structures of personality: it is always social psychology. The system of institutions must be grasped in terms of the imposed repression of needs and of the scope for possible individualization, just as personality structures must be grasped in determinations of the institutional framework and of role qualifications. (Habermas, 1971 [trans]; McCarthy, 1984, p.253)

It is also a matter of empirical investigation, as current research into cognitive variability and linguistic performance indicates (Bialystok & Sharwood-Smith, 1985). It would seem that this broader disciplinary perspective is needed if we are to approach an understanding of an individual's situated significance ascribed to discoursal features. It is tempting to cite Wittgenstein (1958), who ought to be considered the founder of communicative language testing:

> ...but how many kinds of sentence are there? Say, assertion, question and command? There are countless kinds: countless different kinds of use of what we call 'symbols, words, sentences.' And this multiplicity is not something fixed, given once [sic] for all; but new types of language, new language games, as we may say, come into existence, and others become obsolete and get forgotten. (p. 48)

In stressing this first key respect in which my alternative model differs from that offered to applied linguistics (and specifically to testing) hitherto, viz, the introduction into communicative competences of an interpretive dimension focused on capacity and ability, I have concentrated, naturally enough, on meaning-making as a central factor. There are, however, other factors that in principle we ought to be able to evaluate if we are going to get a measure of the communicative competence of our examinee. I refer to the well-attestable variability of nonnative speakers' cooperativeness. Let me explain. Descriptive studies of discourse--for example, the work of Sinclair and Coulthard (1975), Sudnow (1972), Psathas (1979) among many others--all stress the cooperative nature of interaction; people work together to make meanings, much in the manner referred to in this quotation from Erickson:

> Encounters occur within a general social system, and social and cultural influences affect to some extent what happens within the encounters. But encounters also seem to have a life of their own. Persons in encounters are able to make choices among optional specific ways of acting from moment to moment to accomplish those courses of action. Choice is possible among various attributes of status to be attended

> to or ignored. One person's communicative choices from moment to moment constrain the choice of others, and in this sense single individuals are not the sole cause of what happens; social interaction both constitutes and is constituted by the circumstances of enactment. Individuals are part of an ecosystem when they engage one another in interaction. (Erickson 1982, cited in Roberts & Simonot, 1985, p. 181)

We know from these and other studies (see Gumperz, 1982) that nonnative speakers are less able to make use of this discoursal cooperation to achieve their communicative goals than are native speakers (and, heaven knows, there is enough variability among them!). If this is so, then discourse competence presumably becomes a candidate for the evaluation of communicative competence. Moreover, recent studies into the issue of so-called "pragmatic failure" (Candlin, Coleman & Burton 1980; Candlin, 1983; Thomas 1983, 1984) have documented how people are variably able to communicate without trespassing on the personal territories and acceptabilities of their interlocutors, themselves instances of particular views of social reality. Once again, we have another candidate for inclusion in the evaluation of communicative competence.

In short, cooperativeness, acceptability, and capacity--all constructs subject to the influences of personality, ideology, and cognitive ability, and all characterized by varying degrees of conventionality and systematicity--deserve a place in this alternative view of what communicative competence means.

Two current studies, one related to language testing and one not, underscore the points I make.

The first (Roberts & Simonot, 1985), in the context of a Europe-wide project into second language acquisition among adult migrants (see Perdue 1982), explores the effects of <u>context</u> of nonnative speaker (NNS) interlanguage performance and the opportunities afforded for second-language learning to such NNS in noninstitutional settings. The study not only concludes that context is a complex variable (incorporating, for example, what is created as the interaction proceeds, the history of previous interactions, and also the "wider context of living as a member of an ethnic minority group" in Britain), but that our very attempt to reconstruct the context of interaction (as one might do, for example, in a test) involves us in postulating a normative world where presuppositions are allegedly shared. Furthermore, they document how different NNS variably negotiate input, topic choice, and control and variably achieve their short-term and long-term conversational goals. What is significant for evaluation is not the variability, but the manner in which it is dependent on the context of the interaction in question, the context of previous interactions of like kind and, most significantly for a socially responsible testing program, "the socially produced conditions for interaction available to minority members." A similar point is made in Candlin. (1982, 1983)

The second study, now focused on language testing (Douglas & Selinker, 1985) explores the variability in NNS interlanguage as a factor of the "degree of personalization" of the context (in their terms, "discourse domain"). The closer the context of interaction to the familiar worlds of the NNS, the more likely it is to produce interlanguage that is characteristic of the NNS' process competence at that time. Furthermore, the display by NNS of different ability in different communicative contexts is closely related to the degree of engagement of the NNS in the choice and maintenance of topics. They highlight the greater cognitive burden of negotiating unfamiliar topics in (we presume) also unfamiliar contexts, and imply that such conditions require that excessive processing time be devoted to conceptualization rather than to communicative performance. Similar arguments have been advanced by Brown and Yule (1983) concerning variable NNS speech performance under different contextual conditions, and we can readily adduce to the findings of Douglas and Selinker the socially relevant arguments of Roberts and Simonot concerning the communicative opportunities afforded their particular subjects.

It is time to turn to the second key respect in which my view differs from the traditional account. I refer to the concept of <u>explanation</u>. Let us examine some linguistic data:

M: I said within myself 'You know, you don't matter so what are you talking to me for?' And the other one was I felt.
F: What was the sentence 'you don't matter?'
M: I felt I didn't talk directly to you
F: You said some words like, 'you don't matter?'
M: Yes, you don't matter
F: Say this again
M: You don't matter at all
F: Say it again
M: You don't matter at all
F: Say it to a few more people
M: You don't, you don't really matter...

Let us presume, moreover, that this text is one that is presented for interpretation to some NNS speaker, and that we expect him or her to tell us what it means. Immediately, all the indices of an expanded communicative competence come into play and are potentially available as evaluative criteria. The examinee is required to both process the text as object and interpret it as the skilled accomplishment of the two participants. Awareness of linguistic, discoursal, and pragmatic features is called into question, as is the understanding of the particular activity type from which the text is drawn, the transactional structure, and the participants' expected roles. In addition to this ethnographic knowledge, the text makes demands on the examinee's sociological and sociopsychological sensitivity to the frames of reference of both participants and on an evaluation of the quality of "fit" to the encounter of the linguistic features employed. All these necessary demands on communicative competence are open to evaluation if we wish to have a measure of knowledge and capacity.

In this text we might expect an awareness that it was taken from a psychiatric interview between therapist and client, that the purposes of the therapist were X and Y, and certainly that there was a measure of mutual doubt on the precise illocutionary value of the utterances of both parties. All are examples of the variability we have been emphasizing. To this point, then, our examinee is evaluated against criteria of descriptive and interpretive adequacy (though, to be sure, we have not at all specified the nature of these criteria in the world of the text in question, much less associated with them any scalar values). What more is there within communicative competence to evaluate? To answer that, we need to examine further what taking a communicative view of language commits us to. Drawing on the work of Habermas (1970), Berger and Luckman (1966), Brown and Levinson (1976), and Fairclough (1985), it would seem to commit us to at least the following:

1. A concern for social interaction necessarily entails engagement with social theory. (Habermas, 1970)

2. The properties of the social context (especially the power, distance, and impositional relationship between participants) determine discourse. Expressly in Brown and Levinson's words (1978): "...since we see interaction is (a) the expression of social relationships and (b) crucially built out of strategic language use, we identify strategic message construction as the key locus of the interface of language and society." (p. 95)

3. That social structure is not only outside discourse, but within discourse as "shared knowledge" (cf. Berger & Luckman, 1966);

4. That one's choice of social theory is likely to affect how one analyzes (and, we may say, how one evaluates) communication,(cf. Fairclough, 1985).

The requirement of critical explicitness characterizes this second respect. What it does is set an explanatory requirement on both tester and examinee, a requirement to see text as not only shaped by context but shaping it, asking not only what does X mean, or Leech's (1983) questions "what does A mean by X?" but also Haberland and Mey's (1977) question: "How did this utterance and this interpretation come to be produced?"

It is at this point in the argument that an expanded communicative competence as applied to language testing needs my third and reflexive dimension. I refer to the capacity of the examinee to offer an explanatory critique of surface text, to see through it to the substructure of beliefs and social forces. Only if the examinee can do that can he or she hope to explain the particularity of pragmatic interpretations and demonstrate competence. In Ehlich and Rehbein's (1976) terms, "... a functional analysis can only be achieved when one finds the forces and structures underlying socialized life" (p. 65). Furthermore, this explanatory requirement is not only one that relates to the examinee's "background knowledge," it also applies to the interpretive procedures he or she adopts. Indeed, here the reflexive dimension is especially salient. These procedures are also targets for explanation in this expanded view of communicative competence. So far, so good; let us accept that communication patterns are constitutive of a cultural environment as are ways of inferring particular values. In sum, understanding talk requires an understanding of goals, both linguistic and social, and the prototypical notions of what, say, 'doing X' consists of. Moreover, where the frames we "discover" do not have a separate existence from our "background knowledge" itself, as Fairclough (1985) indicates, formed through our particular ideological perspective. Here, however, emerges another difficulty, both for communication and its evaluation: there are social, psychological, and, we would argue, ideological constraints on what is perceived by communicators (and examinees) for interpretation and what is performed, as Douglas and Selinker (1985) specifically point out in the test context, and many experimental studies in language and social psychology (cf Giles & St. Clair, 1979, Ryan & Giles, 1982) amply confirm. Equally, as testers and evaluators, the features we attend to presumably are also affected by knowledge passed to us through our social histories as well as by our particular interpretation of linguistic rules and pragmatic principle. What complicates the issue is that this knowledge and these interpretive constraints are frequently taken for granted, unstated, naturalized, swept under the carpet, and not, apparently, open to question. If we believe this to be so, we ought to attribute a greater and more refined communicative competence to examinees who can critique and explain these hidden presuppositions.

In summary, we ought to value more highly those who can subject the use of particular terms, the choice of lexico-syntactic and phonological realizations, chosen conversational strategies and routines, implied speech act values and interpretive principles, and accepted norms of interaction to analysis and critique. The challenge of this second key respect, like the first, is how to incorporate its implications into our practice of testing.

2. What implications for the limits of testability?

In a trivial and uninteresting sense, there are no limits on testability. Indeed, a thousand years of Judeo-Christian art and the efforts of St. Michael and his angels persuade us of that view. There are, however, real-world constraints, and these constraints have a cost. The question is ultimately economic: do the benefits outweigh the costs, and can we afford to pay?

In attempting this question as a nonspecialist in testing, I am dependent on vicarious expertise derived in particular from the work of my Lancaster colleagues, Charles Alderson (1981) and Pauline Rea (1985). Suggestions that language teaching become more "communicative" (Canale &

Swain, 1980) are by no means novel. It appears that since the mid 1960s, workers have been calling for assessment procedures that would be authentic to target text and task and that would, accordingly, deliver more specific and valid accounts of learners' abilities to understand and convey meaning. Appropriately enough--given the complexity of the subject matter, even in its traditional formulation--research has taken the line that there is not likely to be found any single measure of communicative performance, despite vociferous claims to the contrary, and that, therefore, energies should be directed toward identifying the requirements of any measures proposed, and their concomitant assessment criteria (Alderson & Hughes, 1981). This objective has been made difficult for a number of reasons, some of which I have indicated above. There seems a good deal of uncertainty and pessimism as to whether the content and process characteristics of communication can ever be measured in any summative test.

First, as we have seen, there is a lack of shared assumptions as to what constitutes the nature of communicative competence and what its relationship might be to communicative performance. Indeed, it would be appropriate to ask in this largely post-transformational era whether these divided terms are not as meaningful as they apparently were when Hymes conceived their scope. Certainly, the terms are not sufficiently well defined to offer much hope at present of clear evaluation criteria. Some would argue even that the probabilistic nature of pragmatic value, central to any understanding of communication, is inimical to the traditionally categorical procedures of tests.

Second, notions of authenticity to text and task prove elusive because neither the question of the idiosyncrasy versus the commonality of text structure have been fully explored; nor has there been any such discussion over more than a quite narrow range of activity types.

Third, if the lack of base data under the second point above makes direct connections between test and text/task problematic, we cannot be in any position to indicate what would be appropriate indirect measures of assessment--attractive though it must be to wish to tap communicative competence obliquely.

Fourth, even if such target data were available, it is not immediately clear to me how these could be translated into the content and method of valid criterion-referenced tests (which they would suggest) without extensive pilot testing to check for ambiguity. Indeed, such procedures would inevitably push us toward the establishment of norms, thus begging again the question of authority and control and falling foul of the explanatory criterion discussed earlier in this paper. This is not to say that norm-referenced tests might not be suitable for communicative purposes, only that we do have the data and that there are hidden questions begged.

Fifth, little attention has been paid, as Rea (1985) points out, to the cost-effectiveness of large-scale (traditional) communicative testing--in particular, to determining the trade-off between administrative and evaluative facility and inclusion or exclusion of even those limited communicative possibilities with which I have taken issue.

In essence, the problem appears to lie in determining whether any product-oriented test can deliver that information about processing characteristics associated with good and bad learners (cf Naiman, Frohlich, Stern, & Tedesco, 1975; Rubin, 1981) that we have been arguing lies at the heart of communicative competence--in particular, whether any such test could account for the variability, under a range of linguistic, cognitive, sociopsychological, and sociological constraints, of NNS interlanguage. It is obvious to say that any decision we make will impinge crucially on the issue of authenticity to text, in terms of content validity.

Even if one were to allow that content validity were possible, there will always remain the problem of the degree to which the test procedures themselves mirror the actualities of human communication--whether in terms of skill-use, role variation, variety of purpose, divergences in ideology, or whatever. As Rea (1985) emphasizes, language testing methodology cannot be unaffected by communicative realities. Attention to item types is presumably of equal importance to item content, since, as I have argued, they are just as susceptible to sociocultural influence and personal idiosyncrasy, and affect authenticity to task and domain, in terms of construct validity.

Finally, any test requires concurrent validation and ought not to be seen in isolation from other concurrent assessments of examinee knowledge, capacity, and performance, whether in the academic institution or in the workplace. Indeed, these indirect consumers of test results, with their own opinions on the nature of communicative competence, must have a powerful effect on test design and test validation. Our insistence on a reflexive dimension would confirm their involvement, difficult though that may be for autonomous testers to accommodate.

It is time to pull together what will, I fear, appear as a catalogue of problems that will not easily go away. They constitute challenges to test criteria, whether for TOEFL or any other scheme. They can be usefully set out as follows:

PROBLEMS in establishing the validity of "text" and "task"

PROBLEMS in norm referencing, given the interlanguage variability and the complexity of "context"

PROBLEMS in establishing scales of appropriateness, acceptability, and creative capacity

PROBLEMS in the administration of any process evaluation

PROBLEMS in the evaluation of critical awareness

The answers, I suspect, lie elsewhere than in testing. They depend on improving the quality of our description of communicative competence, and then on making educational and administrative rather than linguistic judgments on where the boundaries of communicative competence are to be drawn. Once we have made our particular peace between these two forces, we can tackle what must be the second order problem, namely the design and validation procedures we advocate as appropriate.

Much discussion in testing is, in my view, vitiated by reversing the order of these priorities. TOEFL is no exception to what is almost a general rule.

References

Alderson, J. C. (1981). Reaction to the Morrow paper. In J. Alderson & A. Hughes (Eds.), Issues in language testing (pp. 45-55). ELT Documents III London: The British Council.

Alderson, J. C., & Hughes, A. (Eds.). (1981). Issues in language testing. ELT Documents III. London: The British Council.

Berger, P. L., & Luckman, T. (1966). The social construction of reality: A treatise in the sociology of knowledge. New York: Doubleday.

Bialystock, E., & Sharwood-Smith, M. (1985). Interlanguage is not a state of mind: An evaluation of the construct for second language acquisition. Applied Linguistics, 1(2), 101-117.

Breen, M., & Candlin, C. N. (1980). The essentials of a communicative curriculum in language teaching. Applied Linguistics, 1, 89-112.

Brown, G., & Yule, G. (1983). Teaching the spoken language. Cambridge: Cambridge University Press.

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

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

Canale, M. (1983) From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), Language and communication (pp. 2-27). London: Longman.

Candlin, C. N., & Breen, M. (1980). Designing and innovating in language teaching materials. Lancaster practical papers in English language education: Volume 2 (pp. 172-216). Lancaster: University of Lancaster.

Candlin, C. N., Coleman, H., & Burton, J. (1980). Dentist-patient communication. Lancaster: University of Lancaster for the General Dental Council.

Candlin, C. N. (1982). English as an international language: Intelligibility versus interpretability. In C. Brumfit (Ed.), English as an international language (pp. 95-98). London: Pergamon Press.

Candlin, C. N. (1983). Beyond description to explanation in cross-cultural discourse. Paper presented at the 1983 East West Center conference on English as an international language: Discourse patterns across cultures. Honolulu.

Douglas, D., & Selinker, L. (1985). Wrestling with context in interlanguage theory. Applied Linguistics, 6(2), 190-204.

Douglas, D., & Selinker, L. (1985). The theory of discourse domains in interlanguage learning. Manuscript.

Ehlich, K., & Rehbein, J. (1976). Sprache im unterricht--linguistische Verfahren und schulische Wirklichkeit. Studium Linguistik, 1, 60-72.

Erickson, F. (1982). The counselor as gatekeeper. New York: Academic Press.

Fairclough, N. (1985). Critical and descriptive goals in discourse analysis. Journal of Pragmatics, 9(6), 739-763.

Giles, H., & St. Clair, R. (1979). Language and social psychology. Oxford: Blackwell.

Gumperz, J. (1982). Discourse strategies. Cambridge: Cambridge University Press.

Gumperz, J. (1984). Communicative competence revisited. In D. Schriffrin (Ed.), Meaning form and use in context: Linguistic applications. Washington, DC: Georgetown University Press.

Haberland, H., & Mey, J. (1977). Editorial: Linguistics and pragmatics. Journal of Pragmatics, 9, 1-12.

Habermas, J. (1970a). Towards a theory of communicative competence. In F. Dreitzel (Ed.), Recent sociology II (pp. 115-148). New York: Macmillan.

Habermas, J. (1970b). Theorie der gesellschaft oder sozial technologie. Frankfurt: Suhrkamp.

Halliday, M. A. K. (1979). Language as social semiotic. London: Arnold.

Hymes, D. (1967). Modes of interaction of language and social life. In J. MacNamara (Ed.), Problems of bilingualism Journal of Social Issues, 23, 8-28.

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

Leech, G. N. (1980). Pragmatics and conversational rhetoric. In G. N. Leech Benjamin (Ed.), Explorations in semantics and pragmatics. Amsterdam: John Benjamins.

Levinson, S. (1976). Activity types in language. Linguistics, 17, 365-399.

McCarthy, T. (1984). The critical theory of Jurgen Habermas. Cambridge: Polity Press.

Naiman, N., Frohlich, M., Stern, H., & Todesco, A. (1975). The good language learner. Toronto: Ontario Institute for Studies in Education.

Oller, J. (1985). Communication theory and testing: What and how? This volume.

Perdue, C. (Ed.). (1982). Second language acquisition by adult immigrants: A field manual. Strasbourg: European Science Foundation.

Pink, J. D. (1982). Observing the teaching of French to a class of slower learners. (Unpublished thesis). Lancaster: University of Lancaster.

Psathas, G. (Ed.). (1979). Everyday language: Studies in ethnomethodology. New York: Irvington.

Rea, P. (1985). Language testing and the communicative language teaching curriculum. In Lee, A. Fok, G. Lord, & Law (Eds.), New directions in English language testing (pp. 15-32). Oxford: Pergamon Press.

Roberts, C., & Simonot, M. (1985). This is my life. How language acquisition is interracially accomplished. (Mimeo) Ealing College of Higher Education.

Rubin, J. (1981). Study of cognitive processes in second language learning. Applied Linguistics, 2(2), 117-131.

Ryan, H., & Giles, H. (Eds.). (1982). Attitudes towards language variation. London: Arnold.

Sinclair, J., & Coulthard, M. (1975). Towards the analysis of discourse: the English of teachers and pupils. London: Oxford University Press.

Sudnow, D. (Ed.). (1972). Studies in social interaction. New York: The Free Press.

Swain, M. (1985). Large-scale communicative testing: a case study. In: Lee, A. Fok, G. Lord, & Law (Eds.), New directions in English language testing (pp. 35-46). Oxford: Pergamon Press.

Thomas, J. (1983). Cross-cultural pragmatic failure. Applied Linguistics, 4(2), 91-112.

Thomas, J. (1984). Cross-cultural discourse as unequal encounter. Applied Linguistics, 4, 226-235.

Widdowson, H. (1983). Learning purpose and language use. London: Oxford University Press.

Wittgenstein, L. (1933, 1958). The blue and brown books. Oxford: Blackwell.

COMMENTS ON THE CANDLIN PAPER

Diane Larsen-Freeman

Although I did not get a chance to see Chris Candlin's paper beforehand, I shall try to make a few comments on it based on what I heard him say here this morning. The fact that I have had only a few minutes to prepare these comments must surely mean that my comments will be an example of a "speeded" response.

Chris made a distinction between conformity to the target language or to conventions (forms) versus the ability to demonstrate creativity (the making of meaning). I see a need for assessing conventions, but I also see a need for assessing creativity, which I would call strategic competence. At one point he gave an example of creativity as being the ability to get on the same wave length with one's interlocutor. I would agree with this, and I believe it is an example of strategic competence in action.

However, I am less optimistic than Chris about the possibility of our being informed by social theory in the assessment of communicative competence. I don't know if we will ever be able to enumerate and agree upon the various indices he suggests. For example, he discussed a speaker's cooperativeness as an evaluative device. It seems to me that when you get into that area, there would be an endless list of indices.

On the other hand, I am fascinated by his idea of an explanatory requirement for the testee; that is, stepping behind the utterance to see if the examinee can explain how a particular interpretation or utterance came to be produced. Of course, one would want to remember that not every native speaker would be able to tell you this, since many native speakers are not aware of why they say what they say.

I have some reservations about his mechanisms for assessing communicative competence. I am not optimistic about his expert system as a mechanism for conducting a process evaluation, but I don't think we should close the door on any idea at this point.

DISCUSSION OF THE SAVIGNON AND CANDLIN PAPERS AND THE RESPONSE BY LARSEN-FREEMAN

Richard Tucker. Before we begin the discussion, I would like to ask Sandra and Chris if they would like to make any reply to Diane's response to their papers.

Sandra Savignon. I wouldn't want to leave you the impression that I don't care about grammar. My concern is rather with the authencity of texts. I think that given the integrative, creative, and compensatory nature of language proficiency, one would not want to pinpoint discrete structural features and say that they are crucial, nor can one say that they can be tested out of context. When one is concerned with authencity of text, then form becomes only a part of meaning.

I am still surprised by the fact that in the early stages of the communicative competence movement, people interpreted me to mean that I advocated a "me Tarzan, you Jane" approach to getting meaning across. Actually, anyone who is concerned with the conveyance of meaning must also be concerned with form. One of the agonies we go through is giving form to the meaning we wish to convey. What I am suggesting is that the focus must be initially on the meaning, and without context there is no language, so you can't measure form isolated from a context.

When talking about nonnative speakers' performance, we should make a distinction between deficiency and deviancy. Forms may deviate from native speaker norms, but I don't think this necessarily indicates a deficiency. To say so implies a value judgement. I don't think one can take a single feature and say "you are deficient in this." Again, I insist on the integrative, creative, and compensatory nature of language proficiency.

Chris Candlin. I would like to thank Diane for making any comments at all on my paper under the circumstances. It think it was very gracious of you to do that.

I don't equate strategic competence with creativity, but regard them as separate, with creativity applying as much to strategic competence as it might to other kinds of competence. We tend to assume that nonnative speakers do not have the right to fail pragmatically, when, in fact, they may deliberately choose to fail, and in doing so indicate their creativity. Native speakers do this also. So the question of norms for assessment is really a difficult one.

I also want to say that it is a common misconception that people who write about communicative competence are not concerned with form. What these people are saying is that form and value are two immensely complicated and interrelated worlds, which we have to treat differently because they are different. It would be as mistaken to treat both in the same way as to suggest that one can exist without the other.

Henry Holtzclaw. I was intrigued with one question that Diane Larsen-Freeman raised. I agree with her that the TOEFL does seem to work in that it gives us a view as to whether a person can handle graduate study, at least from a linguistic standpoint. And I was intrigued by the reverse of that prediction. That is, does a less-than-satisfactory score on the TOEFL guarantee that a person cannot handle graduate study? I am not sure how one could research this, unless a group of administrators from across the country were to admit students with less-than-satisfactory scores. We do that sometimes, but not in so systematic a way as to draw conclusions from it.

Diane Larsen-Freeman. One way to do such a study is to identify such students and administer alternative (communicative competence) instruments to them in addition to the TOEFL and to relate student performance on them to performance in school. If, by comparison, the TOEFL were not the best measure, the institutions would be free to require applicants to take a different test. At present, students who do poorly on TOEFL have no other alternative than to take the test.

Dan Douglas. I would like to comment on Diane's suggestion for alternative instruments. Many years ago the University of Michigan tried to make practical statements about what the examinee can do at different score levels and to adjust these by field and udergraduate or graduate candidate status. I don't think anyone paid attention to it, but they did try to do it. It may be possible for TOEFL to serve as a "broker" of interpretive information in a similar way, as well as by relating information on performance on alternative types of tests, and by distributing examples of examinee performance on productive skills tests.

Frances Hinofotis. We should remember that the TOEFL program emphasizes that the TOEFL should not be the only measure or indication of competence on which the admissions decision is made.

Lyle Bachman. If you look at predictive validity studies of TOEFL using academic achievement as a criterion, the results are usually quite miserable. But I might also add that there are not any tests that are very good predictors of academic achievement. One of the reasons for its poor predictive validity is due to the fact that language proficiency is not the critical factor in academic achievement. The reason TOEFL works is partly because it is the only show in town. As the years have passed, institutions have gained considerable experience with TOEFL and have adjusted their TOEFL score requirements to the point that they feel comfortable that they are making the right decision if a person scores above their requirement.

Diane Larsen-Freeman. Why do institutions feel comfortable about TOEFL? It may be because it does more than satisfy a need. We know it measures some aspect of language proficiency. It probably correlates fairly well with communicative competence.

Lyle Bachman. In my comment I didn't mean to say that TOEFL doesn't work. My point is that it works because it is institutionalized, but that doesn't mean it measures something that is relevant. We might give a mathematics test and find that it works equally well. That is one of the problems with looking at predictive validity; you can make no assumptions about content or construct validity just because something works.

Russell Campbell. At UCLA we test about 1,200 to 1,400 foreign students a year with our ESL placement test. The results indicate that many of these students, instead of studying physics, chemistry, engineering, and so forth, should be spending about two-thirds of their time studying English. But instead of doing this, they take other courses. Quite often, when we examine their academic records after three of four years of study, we find that they have high grade point averages. So people with major English language deficiencies may do very well in their academic programs. They do this by using various compensatory strategies. They tape record the professor's lecture and play it back during the evening, and they spend up to five times the usual amount of time reading an assignment.

John Oller. At a meeting designed to help us prepare for this conference held at the 1984 TESOL International Convention in Houston, Chris Candlin suggested that a criterion for determining the validity of the TOEFL ought to be its success as a predictor of academic success. It reminded me of one of my earlier experiences as a junior faculty member at UCLA. At that time Clifford Prator put me in charge of the ESL Placement Examination. In that role one day I mentioned to him that one way to validate our test would be to examine its ability to predict academic success. Prator explained to me that if the placement exam predicted ultimate success in school, then our English language program either would be a failure or it would be unnecessary. Similarly, if students improved their scores on the placement test after taking our courses and then subsequently did well in school, that would negate the predictive validity of the earlier score. What I conclude from this is that incoming English proficiency is not directly related to success in a university. There are so many other factors that play a role that it is difficult to attribute either academic success or failure to English language proficiency. This being the case, we can not expect a language proficiency test to predict academic success.

A second point I would like to make relates to why some of us feel comfortable in saying that the TOEFL measures English language proficiency. We know that TOEFL is a fairly good indicator of grammatical skills and knowledge. We also know it is a fairly good indicator of reading comprehension and vocabulary. We know a little bit less about how well it measures strategic and discourse competence and creativity in the sense that Dr. Candlin uses the term. Still, we can be fairly confident that it does measure English language proficiency fairly well. In fact, to the extent that we know what communicative competence is, we are forced to conclude that TOEFL does a fairly good job of measuring it.

Carlos Yorio. I would like to go back to the question of grammatical competence. Based on my experience working at an urban university, it seems that mere mutual intelligibility alone is not sufficient to be a successful communicator. I find that flaws in form are highly stigmatizing to students in my program, and in fact, that is why many of them are coming to the program after having lived in this country for between five and twenty years. I offer as an example the case of my secretary. Many of my colleagues in the university refuse to leave a message for me when they call and find that I am not in. When I ask them why, they tell me that my secretary would not understand them. I am certain she can be understood and that she could understand the caller, yet people refuse to negotiate meaning with her. Thus, I think we ought to be more sensible about grammatical competence, given its sociolinguistic effects.

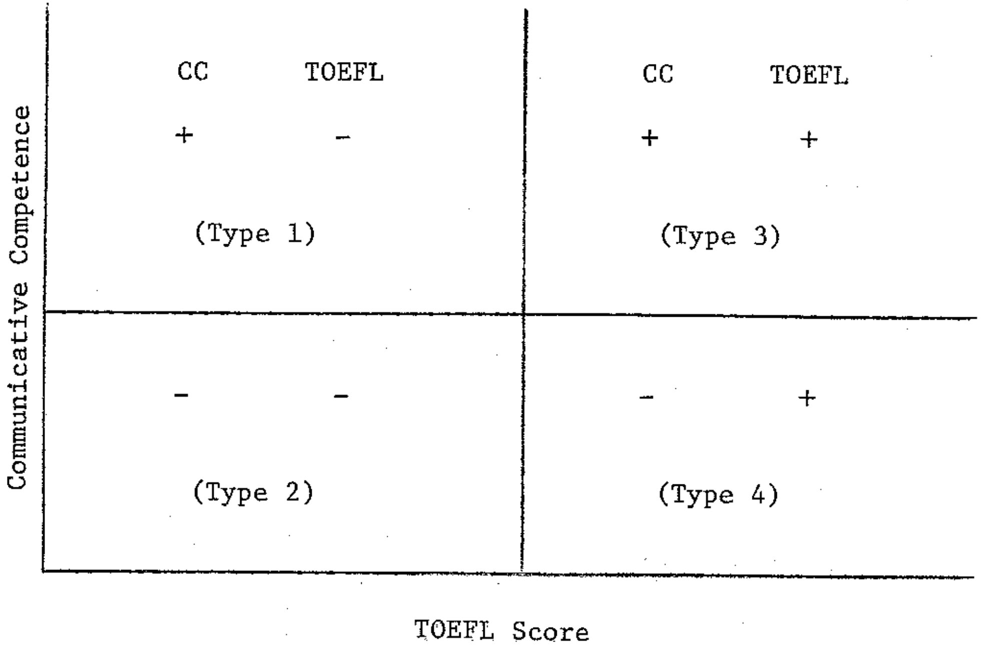
Kathleen Bailey. If we consider that TOEFL is measuring something and that communicative competence is potentially something else that could be measured, and if we consider the TOEFL requirement of a university admissions office as the dividing line between an admit-score and a reject-score, then we can we depict these variables as indicated in the following figure. We can talk about people who are communicatively competent but do not score well on the TOEFL. This would be student type 1. We can also talk about people who are not communicatively competent and do not score well on the TOEFL; that would be student type 2. Similarly, in the right quadrants we have type 3, the student who is communicatively competent and scores well on the TOEFL, and type 4, the student who is not communicatively competent but scores well on TOEFL.

I think this demonstrates some interesting things to us as researchers. First, we must recognize that students in the left half of the figure are not available to us as subjects since they are not admitted to the university, unless they repeat the TOEFL, as Ken Wilson's (1986) study suggests they do, and then move into another quadrant. Yet if we can design a measure of communicative competence, we would want to gather data on people who are in the first quadrant. We need to know a lot more about this type of individual. Of course, I don't suggest that we admit, as Henry Holtzclaw intimated, large numbers of people and then see if they sink or swim. But we do need to conduct research on this type of ESL learner.

Lyle Bachman. We could also think of Kathi's figure in terms of the admissions decision. If we replace communicative competence with academic success, we see each admissions decisions has a cost associated with it. In quadrant 4 we have students who were admitted with acceptable TOEFL scores and yet did not do well in school, and in quadrant 1 we have students who were not admitted yet who would have succeeded had they been admitted. Thus, there are costs associated with both of these groups. If we accept a student on the basis of an acceptable TOEFL score and then learn that the student needs further ESL training, it costs the institution, which has to provide the instruction, or it may cost the student, who has to spend five times as much time on reading assignments as students who are adequately trained in English. I think even more insidious is the cost of rejecting students who ought to be here. Most of these students are the best that their countries have to offer. Thus,

Figure 1

Interrelationships between Communicative Competence and TOEFL Scores



their rejection has tremendous social costs to the countries of origin. It gets back to Kathi's point. We need data on the people who score poorly on TOEFL and yet would otherwise be successful. Unfortunately, at present we only see the cost of accepting people who turn out to be inadequate. We don't see the cost of rejecting people who could have been successful.

Henry Holtzclaw. At the University of Nebraska, in cases where the student looks very promising in all respects except TOEFL score, we do allow the professor to call the applicant and talk with him or her. So not all these students are lost to us, and it may be possible to identify similar students at other universities.

Charles Stansfield. There have been a lot of studies of the predictive validity of TOEFL. Recently, my colleagues and I conducted a comprehensive review of the literature on TOEFL (Hale, Stansfield, & Duran, 1985). We found studies showing a correlation with grade point average (GPA) ranging between -.40 and +.80. In the case of those two extremes, the samples involved fewer than ten people.

The best study we identified was one conducted by the American Association of Collegiate Registrars and Admissions Officers (see pp. 27-30). It involved over 1,000 foreign students who had received scholarships from the Agency for International Development. The students attended a large number of American universities and studied a wide range of academic disciplines. The study compared the relationship between their GPAs and scores on the TOEFL and the ALIGU (American Language Institute of Georgetown University) tests. The correlation between GPA and TOEFL was .25, and the correlation between GPA and ALIGU was .22. However, when the authors correlated the two tests with both GPA and number of credits earned during the first year of study, the correlation went up to .36 for TOEFL and .32 for ALIGU. Still, even these latter correlations are low, and I suspect this is due to a number of reasons.

First, among people who are highly proficient, there is not going to be any relationship between TOEFL score and academic performance. In fact, native speakers average about 630 on the TOEFL, yet there is no relationship between being a native speaker and success in college. And so as second language learners reach the native speaker level, I suspect their academic performance would begin to behave in the same way.

Second, this is probably due to the fact that language proficiency acts as a threshhold variable in its affect on academic achievement. That is, once a learner reaches a certain level of level of language proficiency, language is no longer an obstacle to academic success. Other factors, such as academic aptitude, prior training in the field, and "personological" variables, then enter into the picture and play the principal role in determining academic performance.

We don't really know what the threshhold level is. Certainly it is no higher than the native speaker level (630), and I suspect it is a good deal lower, probably lower than 600. And since the average institution doesn't admit applicants with scores lower than 550, we would not expect to see any relationship between TOEFL score and success for a good number of the foreign students who get admitted. As has been mentioned before, we don't get to see the people at the lower end of the TOEFL distribution. And unfortunately, because we don't have research findings on them, not much has been done to look at what measurement alternatives may exist for these people. We don't know if these applicants are being treated appropriately at the moment.

Graduate institutions frequently do admit students based on the recommendation of a professor, even though the individuals do not fulfill the requirements for admission. What we really need are studies that would look at these people in terms of their personal characteristics and the kind of language measures they can perform well on and see how these variables relate to their academic success. When we have this information, it may be possible to incorporate it into the TOEFL program. So I would encourage those who work at such institutions to conduct research on these individuals.

Harold Madsen. I would like to return to Chris Candlin's paper and his earlier remarks about communicative competence and put a question to him. What do you see as the practical application of your view of communicative competence to the TOEFL?

Chris Candlin. Let's imagine that we have a lady who is nonnative speaker of English and that she wishes to obtain employment as a family planning counselor in a multi-ethnic community. We have to find a way of assessing her competence to perform this task in English. We have two choices. We could give her a test that just asks questions about grammatical models or we could step beyond that. If we do go beyond a grammatical model, a number of issues arise.

Let's say that that we could confront the would-be counselor with a potential client who is both pregnant and a chain-smoker. The counselor knows that she is supposed to give information, not advice. But, the client's situation involves two contraindicated conditions. So the counselor must try to use her strategic competence to present the information in such a way that the client is motivated to stop smoking, but without giving advice. This is what I want to assess--the ability to function under the ideological basis of a family planning center. When you move beyond strict assessment of grammatical skills, you must get into the area of domain specificity. Now I don't think this can be accommodated in the TOEFL. However, if you want to test communicative competence, you must develop other measures in addition to TOEFL that will assess the examinee's ability to function in a specific domain.

Harold Madsen. I am concerned about this matter of domain insofar as it relates to developing the test. It seems that any tests that were developed would be applicable to a very limited group of people.

John Oller. It seems to me that what most of us who in recent years have argued for pragmatic/integrative tests are saying is that what the TOEFL now measures, and what many other integrative and pragmatic tests measure, already is, in some fundamental and indisputable sense, communicative competence. To suggest that what we need to do is something radically different is an extraordinary misinterpretation of what is going on in the field. It has not been demonstrated, to my knowledge, that the TOEFL measures something fundamentally different from what these other tests measure. The idea that one can define a strategic competence that is something totally different from and independent of what is being measured on pragmative/integrative tests is a kind of heresy. The evidence shows that is it difficult not to measure communicative competence whenever you have people communicate. It seems to me that the critical issue is whether tests in the educational domain, like the TOEFL, involve a kind of communicative competence. I don't think that there is any fundamental aspect of communicative competence that is not measured by the kind of textual reasoning that is required to intereact with the text currently in the TOEFL. I disagree with Professor Candlin in that I don't see room for an alternative model. I don't see any distinction in terms of theory between what we are proposing to do and what we are in fact already doing. To say that integrative tests don't count as communicative tests is to miss the whole point of the discrete point/integrative controversy that John Carroll brought up twenty years ago. The whole of integrative test literature asserts that what we are measuring with integrative/pragmatic tests is in some fundamental sense communicative competence.

Lyle Bachman. John has brought out the idea that the fact that all these tests are related to each other must suggest that they are all measuring communicative competence to some degree. And Diane has suggested that as an experiment we give other tests to people who do poorly on TOEFL and see how the scores correlate. The fact that all these tests are correlated with each other doesn't answer the basic question of what they test. A few years ago the Foreign Service Institute (FSI) was interested in looking at the validity of the FSI Interview. They brought some people together to talk about how this could be done. The kinds of things people talked about were performance tests. Someone suggested that examinees be invited to dinner and that a microphone be planted under the table. Yet that's not any more of a performance test than the FSI Interview. So we are just going around in circles when we propose trying to validate TOEFL on other measures. The other measures will be susceptible to the same kinds of limitations as TOEFL. They are the limitations in performance, in context, and in domain that we are already attributing to TOEFL.

Kathleen Bailey. Chris seemed to be calling for situation-based performance tests. And yet what I heard Lyle say was that language proficiency can't be measured by performance tests, or was it criterion-referenced tests? What did you mean?

Lyle Bachman. What I was saying is that to treat a performance test as a criterion-referenced test is a mistake. With a criterion-referenced test you start with a definition of a domain or a criterion. In order for Chris to have a criterion-referenced test, he would have to be able to describe the parts of the domain. He talks about illocutionary force, about strategic competence, and oral context, and these are all parts of a domain. But a performance test can't really be treated as a criterion-referenced test until we can define the domain. Second, we can't treat a performance test as a criterion-referenced test until we can define the scale that would be used to assess performance. We would need to know what a zero on the scale would be or what a perfect performance would be. If people can take different routes to accomplish a task successfully, then we can't specify a criterion. I think performance tests are simply more detailed, more extensive samples of behavior. And when we talk about samples of behavior, we are talking about norm-referenced testing. I am not uncomfortable with the fact that we are talking about norm-referenced testing. The fact that we establish a desirable score level on a test like the TOEFL does not make it a criterion-referenced test. In order to have a criterion-referenced test, the facts must be fixed, and in real-life communication, the facts are hardly ever fixed.

Frances Hinofotis. I think there may be some confusion about the difference between norm-referenced and criterion-referenced tests. With a norm-referenced test, you are comparing one person against another. With a criterion-referenced test, you are comparing an individual against a standard of performance. I have no concerns about an institution requiring a 550 on the TOEFL so long as it can demonstrate with research that the score level is appropriate. From a test development standpoint, we must recognize that frequently the two approaches are mixed.

Carlos Yorio. Everything we have said thus far in terms of communicative competence applies to foreign languages and foreign students, but I think that it does not apply to urban bilingual students who were born in this country or have lived here a long time. These students are communicatively competent, but yet many of them speak a pidginized English.

Lyle Bachman. I think this is what Cummins (1980) talks about in his distinction between basic interpersonal communication skills and cognitive academic language proficiency. These are two types of language-use situations. We might examine TOEFL from the viewpoint of the extent to which it taps those two situations. I think that if TOEFL errs on one side, it would be on the side of cognitive-academic language. There may be some components of communicative competence that we have talked about that are less relevant to the academic setting than others. For example, certain elements of sociolinguistic competence may be less important than certain aspects of discourse competence.

References

Cummins, J. (1980). The cross lingual dimensions of language proficiency: Implications for bilingual education and the optional age issue. _TESOL Quarterly_, _16_(2), 175-187.

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 17). Princeton, NJ: Educational Testing Service.

Wilson, K. E. (1986). _Patterns of test taking and score change for examinees who repeat TOEFL_ (Final report submitted to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

THE TEST OF ENGLISH AS A FOREIGN LANGUAGE AS A MEASURE OF COMMUNICATIVE COMPETENCE

Lyle F. Bachman

INTRODUCTION

Since its inception in 1963, the Test of English as a Foreign Language has become the _de facto_ operational definition of the construct "English language proficiency" for thousands of institutions of higher education in North America and consequently for hundreds of thousands of nonnative English speaking applicants to these institutions. In these twenty years, our conception of language proficiency and how it is acquired as well as the methods and techniques for teaching it has been greatly expanded. Perhaps the distinguishing characteristic of this expanded conception is its recognition of the importance of context beyond the sentence to the appropriate use of language. This context includes both the discourse of which individual sentences are a part and the sociolinguistic situation that governs, to a large extent, the nature of that discourse, in both form and function.

As Upshur (1971) noted several years ago, language testing practice generally lags behind theories and methods in language teaching. And the TOEFL is no exception. In the twenty years since its initial development, the TOEFL has undergone virtually no major revisions, either in content or format (other than the restructuring, in 1976, from five to three parts). The theory of language proficiency upon which the TOEFL was originally based was quite compatible with an operational definition that permitted the testing of separate, "discrete points" of the language in distinct, unrelated contexts that were generally limited to a single sentence or two. This operational definition was also well suited to current psychometric models and the limitations they imposed. The TOEFL was an example par excellence of the application of the most advanced theories of language abilities and measurement to practical test development.

As noted above, current theories of language proficiency are much more inclusive than that upon which the TOEFL was based, and there would seem to be little reason to expect that the TOEFL would tap the different competencies that are part of these theories. For this reason, it is commendable that the TOEFL Policy Council has requested that the test be scrutinized and evaluated in terms of this particular set of constructs, which it was not explicitly designed to measure. The results of such an examination have implications for substantive revisions in both the content and format of the TOEFL. And such revisions, which could include greater contextualization and variety of test tasks and require the processing of extended segments of discourse, have major implications for current psychometric models, which generally assume unidimensionality of scales and local independence of test tasks.

In this paper I will first attempt to clarify some terms and concepts that will be of use in examining the TOEFL within a current theoretical framework of language proficiency. I will then examine, within this framework, the nature of the performance tasks and the language competencies required by the TOEFL. Finally, I will discuss the implications that this examination suggests for psychometric theory, for revisions in the TOEFL, and for future TOEFL-related research.

A FRAMEWORK FOR A TERMINOLOGY OF LANGUAGE PROFICIENCY

Since many of the terms relating to language proficiency have been used in widely differing ways, I would like to summarize a framework that Adrian Palmer and I have proposed for describing performance on language tests and the different factors that affect language test scores (Bachman & Palmer, 1984, and forthcoming). The factors that affect language test scores are summarized below.

FACTORS AFFECTING LANGUAGE TEST SCORES

Trait Factors (Competencies)

- Organizational
    - Grammatical (Lexis, Morphology, Syntax)
    - Discourse (Cohesion, Rhetorical Organization)
- Pragmatic
    - Illocutionary (Language Functions)
    - Sociolinguistic (Register, Dialect, Figurative Language and Cultural Allusions, Naturalness)

Skill Factors

- Psychophysiological (Mode: Productive/Receptive, Channel: Auditory-oral/Visual)
- Forms of Representation (Conscious/Unconscious, Prefabricated Patterns, Prefabricated Routines, Rules)
- Strategic Competence

Method Factors

- Language-Use Situation (Reciprocal/Nonreciprocal, Transitive/Reflexive)
- Amount of Context (Embedded, Reduced)
- Distribution of Information (Compact, Diffuse)
- Type of Information (Concrete, Abstract)
- Artificial Restrictions (Organization of Discourse, Propositional Content, Illocutionary Force, Forms, Participants, Mode, Channel, Time/Length)

Random Factors

- Affective Factors (Personality, Cognitive, Motivational)
- Interactions among Trait, Skill, and Method Factors
- Measurement Error

Trait factors are those competencies or mental abilities that are related to language use. These are of two main types: organizational, consisting of grammatical and discourse competence, which pertain to the formal characteristics of language usage, and pragmatic, consisting of illocutionary and sociolinguistic competence, which pertain to the functional characteristics of language use.

Skill factors are those characteristics of the individual that affect test performance. These consist of (1) psychophysiological skills, which are distinguished in terms of mode (productive/receptive) and channel (aural-oral/visual), (2) forms of representation (conscious/ subconscious, prefabricated routines and patterns, rules), which determine the extent to which language competencies are available for use, and (3) strategic competence, which affects how language competencies are utilized for maximum effectiveness in processing information.

Method factors are those characteristics of the test method that affect performance. These factors are distinguished in terms of (1) the type of language use situation (reciprocal/nonreciprocal, transitive/ reflexive), (2) the amount of context, the distribution and type of information presented, and (3) the type and degree of artificial restrictions on the language performance required. Communicative methods are those that involve transitive, relatively unrestricted, appropriately contextualized language performance, while noncommunicative methods involve only reflexive, artificially restricted, inadequately contextualized language performance.

Finally, random factors consist of (1) affective factors in the individual, (2) interactions among specific combinations of trait, skill, and method factors, which may vary across individuals, and (3) measurement error.

We would propose to use this framework for defining various aspects of language proficiency in terms of the portion of variance they contribute to test scores.

> Linguistic competence consists of the trait factors of grammatical competence, which pertains to the formal organization of language: syntax, morphology, phonology, and graphology.
>
> Communicative competence consists of linguistic competence <u>plus</u> those trait factors that pertain to the organization of discourse (cohesion, rhetorical organization), the performance of speech acts, or language functions, and the sociolinguistic conventions of appropriateness.
>
> Language skills (listening, speaking, reading, writing) consist of trait and skill factors.

> Linguistic performance consists of linguistic competence, skill factors, and noncommunicative method factors.
>
> Communicative performance consists of communicative competence, skill factors, and communicative method factors.
>
> A measure of linguistic performance includes that portion of test scores attributable to linguistic competence, skill factors, noncommunicative method factors, and random factors.
>
> A measure of communicative performance includes that portion of test scores attributable to communicative competence, skill factors, communicative method factors, and random factors.

The above framework can also help clarify some misconceptions regarding the terms "direct" and "indirect" as they have been applied to language tests. The term "direct test" is often used to refer to a test method in which performance resembles "actual" or "normal" language performance, while an "indirect test" is one in which test performance is perceived as somehow different from "actual" or "normal" performance. Thus, writing samples and oral interviews are referred to as "direct" tests, since they presumably involve the use of the skills being tested. By extension, such tests are often regarded, virtually without question, as construct valid and, therefore, legitimate criteria for the validation of "indirect" tests.

There are two problems with this, however. First, we have no definition of "actual" or "normal" language use that is precise enough to determine whether performance on a given test is or is not similar to such language use. The framework suggested here may at best permit us only to distinguish relatively "communicative" from relatively "noncommunicative" language performance. A more serious problem is that the use of the term "direct" confuses the behavioral manifestation of a trait or competence for the construct itself. As with all mental measures, language tests are indirect indicators of the underlying traits in which we are interested. The framework presented above captures this distinction by recognizing that there are factors in addition to trait factors that affect performance on all language tests, regardless of whether these require recognition of the correct alternative in a multiple-choice format or the writing of an essay, or whether they are "discrete-point" or "integrative."

AN EXAMINATION OF THE TASKS REQUIRED AND COMPETENCIES ASSESSED BY THE TOEFL

In examining the TOEFL as a measure of communicative performance, I will address two questions:

(1) To what extent do the tasks required on the TOEFL involve communicative language performance?

(2) To what extent does the TOEFL assess communicative competencies?

I will attempt to address these questions by examining the same form of the TOEFL (3FATF5) that Duran, Canale, Penfield, Stansfield, and Liskin-Gasparro (1985) examined. I view this examination in much the same light as they view their study: as an initial attempt rather than a definitive description. This examination is subject to the limitation that it is based on a single form of the TOEFL. An additional limitation is that I have not had access to the TOEFL table of specifications and so will not attempt to judge whether or not the performance tasks and competencies I have identified are those intended by the item writers.

I feel that this examination is complementary to the Duran et al. study, in that, while their objective has been to provide an overall description of the language content of the TOEFL, I will focus only on the nature of the performance tasks required and the competencies necessary to successfully complete these tasks. The discussion that follows describes the tasks required by TOEFL items in terms of four characteristics: (1) the restrictions imposed by the task, (2) the specific nature of the task itself, (3) the competencies required to successfully complete the task, and (4) the role of context on the performance of the task.

The language performance tasks on this form of the TOEFL are all restricted to nonreciprocal situations in which there is no potential for feedback or negotiation of meaning. In addition, it would appear that virtually all the tasks are primarily reflexive, involving the demonstration of the test taker's competence. It is possible that some of the tasks, particularly in the reading comprehension section, may be perceived by the test taker as performing authentic illocutionary acts. To the extent that this is so, these tasks involve transitive language use. The format of the TOEFL as a whole restricts the mode of performance to reception. Finally, there are obvious restrictions on time and length.

Section 1, Listening Comprehension

In general, the language performance in this section of the test is restricted to academic and general topics in terms of propositional content, it is ideational and manipulative (instrumental and regulatory) in terms of illocutionary acts, and the channel is restricted to audio. The participants include hypothetical speakers who deliver single sentence pronouncements, a lecture, and an announcement, and of a cast of hypothetical characters who interact with each other in a variety of ways. A wide range of specific tasks is required. The majority of the items tap only grammatical competence, with a smaller number also tapping discourse competence. Only one item, it seems, taps sociolinguistic competence.

Part A, Statements

Restrictions:

Only single sentences are included in this part, and the forms are restricted to declarative and imperative sentences.

Tasks:

There are two basic tasks required in this part: (1) to comprehend a single spoken sentence (stem) and (2) to recognize the correct paraphrase (key) of this sentence from among four written sentences. The nature of the relationship between the stem and the key response varies and may involve primarily lexical synonymy (e.g., item 1), syntactic transformation and lexical synonymy (4), the reversal of a construction and lexical antonymy (7), inference involving real-world knowledge (11), or similarity in illocutionary force (8). This variation may well account for the majority of the variation in difficulty level of these items.

Competencies required:

The majority of the items in this part require only grammatical competence for successful completion. Many of the items depend primarily on knowledge of lexical signification (e.g., items 3 and 5) or of propositional content expressed by syntactic structure (item 15), or both (items 9 and 16-20). Only two (items 4 and 11) appear to require knowledge of cohesion, and only two (items 1 and 8) appear to involve knowledge of language functions. None of the items appears to tap sociolinguistic competence.

Characteristics of context:

Virtually all of these items can be regarded as context-reduced, in that they are completely unconnected with each other and their references are to persons, objects, and actions that are unknown to the test taker. In item 15, for example, the listener does not know who Dorothy is, what she could have helped him or her with, or why he or she is interested in finishing whatever in a hurry.

In general, this part of the test would appear not to require communicative performance. The task of recognizing paraphrases requires virtually no communicative performance, since the paraphrase merely repeats the information in the statement. In addition, the lack of context, as well as the general restrictions mentioned above, renders this part highly artificial. Finally, it would appear to tap primarily grammatical competence.

Part B, Dialogues

Restrictions:

Discourse organization is restricted to single conversational exchanges, while there are various speech acts, including ideational (informative, expressive) and manipulative (request, regulatory, interactional). Forms are restricted to declarative and interrogative sentences.

Tasks:

The primary tasks in this part involve comprehending a spoken dialogue and then answering direct information questions based on the dialogue. The type of information requested in these questions varies a great deal and includes literal propositional (item 24), presuppositional (item 21), and inferential (item 27) information. The specificity of the questions varies as well, from the very specific (item 31, "What does the woman advise the man to do?") to the quite vague (item 30, "What does the man mean?"). It would appear that this variation in type of information requested and specificity of question may be one source of variation in difficulty among these items.

Competencies required:

Nearly half of these items appear to require only grammatical competence. Item 24, for example, requires only the knowledge of lexical signification and propositional content expressed by syntactic structure to answer correctly. Three items (21, 22, and 27) appear to tap knowledge of cohesion as well.

There are four items (21, 28, 30 and 35) for which knowledge of illocutionary acts may either facilitate the correct answer or make the item more difficult. If the test taker recognizes the relationship-defining function of the greeting in the dialogue in item 21, for example, this may help him or her answer correctly. In items 28, 30 and 35, however, the illocutionary force of the statements is ambiguous. In item 30, for example, while the woman says, "The map shows that this street goes downtown," the illocutionary force may be expressed more directly as "I want to go downtown." Likewise, the man's response, "Yes, but what we want to know is how to get to the park," could also be interpreted to mean "I want to go to the park." Given this interpretation, choice (B), "He wants to go to the park, but she doesn't," might be considered an acceptable answer. And what about the man, in item 31, who is diplomatically asking the woman to type his paper and gets turned down, just as diplomatically? Or the man in item 35 whose attempt at "culture" is rebuffed? Does the woman "mean" that (C) "She thinks the book is excellent?"

Two items (21 and 23) presuppose knowledge of information outside the discourse itself. To answer item 21 correctly, the test taker must know where one goes to buy traveler's checks and open a savings account, while in item 23 he must know that, in countries with winter weather, the ground becomes covered only after "it has been snowing for some time." One item (27) requires not only extratextual knowledge, but also inference. To correctly answer this item, one must know that a bicycle has one front tire and a seat that can be raised and that a flat tire needs to be repaired. Items such as these, which require the test taker to identify and recall the relevant information from outside the discourse, may well tap strategic competence, to the extent that this ability is associated with general information processing.

Finally, one item (32) appears to tap one aspect of sociolinguistic competence: the knowledge of figurative language, in the expression "goes in one ear and out the other."

Characteristics of context:

Although there is more context within each item in this part than there is in Part A, the items in this part must still be regarded as largely context-reduced in the same way: We do not know who these people are or what they are referring to and can only guess what their relationship is to each other.

In summary, this part of the test does not seem to require communicative performance, but it does seem to tap a wider range of competencies than does Part A. While many items only require grammatical competence, several also tap discourse and illocutionary and strategic competence; one taps an aspect of sociolinguistic competence.

Part C, Short Talks

Restrictions:

There is a variety of discourse organization, including generalization and development, conversational exchanges, and announcement. There is also a variety of illocutionary acts and forms.

Tasks:

The basic tasks in this part are (1) to comprehend a spoken discourse and (2) to answer direct information questions based on that discourse. In this part the extent of the discourse, at least in the first two texts, is much more substantial than that in Part B. The type of information requested is varied, although the majority of the items request only literal, propositional information (e.g., items 37, 43-46, 48-50). Some require inference or "real-world" knowledge as well (items 36 and 47).

Competencies required:

The majority of the items appear to assess only grammatical competence. Item 48, for example, requires only the lexical knowledge that the phrase "in a few minutes" signifies "very shortly," while item 37 requires equating the lexical significations of "large deaf population" and "a large number of its residents are deaf." Four items (36, 38, 40, and 42) appear also to tap knowledge of cohesion. Item 38, for example, requires a recognition of the cohesive ties between "They," "residents of the island," and "an excellent example of ... inheritance of deafness." Item 36 appears to tap illocutionary as well as grammatical and discourse competence, since the recognition of the relationship-maintaining function in the first sentence may facilitate recognition of the correct response. Finally, item 47 appears to tap strategic competence, in that it requires the test taker to recall the relevant information that library employees would ignore this announcement, since they would already know when the library closes, and that students use reference and reserve books for research.

Characteristics of context:

The context in this part is generally much more extensive than that in Parts A and B. This is not to say, however, that it is necessarily more conducive to communicative performance. The lecture is highly artificial in several respects: the rather gratuitous greeting, the passing reference to a reading assignment (two books!), and the "read" presentation. The dialogue that comprises the second text is even more artificial. First, it lacks coherence in at least two places ("You said that Muir Woods is near San Francisco?" and "I've heard that many redwood trees . . . ."). The woman apparently either is not really interested in what the man has to say or enjoys breaking the thread of conversation. The net effect of this artificiality is to make the task more difficult than it might be were the texts more authentic. There is little challenge to the test-taker to interact with the text, and consequently less information may be processed. In this respect, I concur with Oller's assessment of this section (this conference).

This section as a whole has much potential for involving communicative performance. Despite the restrictions on language use situation, participants, mode, channel, and time imposed by administrative and psychometric constraints, I believe communicative performance, or authentic language use, could be achieved by following some of the suggestions made by Oller (this conference). I see no reason why the inclusion of hesitations, false starts, rephrasings, and other characteristics of "natural" communicative language use would in any way compromise the acoustic fidelity of the recording.

In addition, I would consider eliminating Part A entirely. It has little potential for authenticity and taps competencies that are quite adequately measured in other parts of the TOEFL. Also, since it is the first part of the test, it may serve to reinforce any preconceptions of artificiality test-takers may already have regarding the TOEFL.

Section 2, Structure and Written Expression

The language performance in this section is restricted entirely to single sentences and thus involves no discourse. There is a variety of propositional content, including topics from the sciences, agriculture, and history as well as some general topics. Virtually all the items are restricted to the ideational function, and the form used is restricted to declarative sentences. The participants are a hypothetical "academic" author and the test taker. The channel is restricted to the visual. By far the majority of the items require only grammatical competence, as would be expected for this section, although a few would appear to require discourse and illocutionary competence as well.

Items 1-15, Structure

Task:

In this part, the basic task is to recognize the syntactic form that will correctly complete an incomplete statement. In most cases, it would appear not to be necessary to process all the information in the stem to correctly answer the item.

Competencies required:

The items in this part require syntactic competence almost exclusively. Two items (14 and 15) may involve other competencies as well. In item 14, distractor (B), "been long symbols," is grammatical, but lacks naturalness, and means something different from the key, "long been symbols." In item 15, distractors (A), "the past," and (D), "those past," are also grammatically correct. What makes them incorrect, I believe, is that they do not make as clear a cohesive reference to "libraries of the past" as does the key, "those of the past."

Characteristics of context:

The items in this part must be considered largely context-reduced, in that they represent isolated propositions, although there has been some attempt to contextualize them. In by far the majority of the items, however, this context is totally irrelevant to the task posed by the item. Consider item 1, for example:

> Conifers first appeared on the Earth _____ the early Permian period, some 270 million years ago.
>
> (A) when
> (B) or
> (C) and
> (D) during

The statement, as found in a reading text, would presuppose that the reader is familiar with the terms "conifers" and "Permian period." Although this information has little to do with the syntactic structure that requires the preposition "during," if the test taker attempts to process this sentence as an authentic use of language and is not familiar with these terms, the item becomes context-reduced and may be more difficult. Other difficult-to-process items in which the context is largely irrelevant to the task required are items 4 and 13. In these items the terms may be unfamiliar and the information is quite compact. Because of this unnecessarily difficult context, these items probably engage other competencies, even though they are intended only to measure grammatical competence.

Items 16-40, Written Expression

Task:

The task posed by this part involves essentially the reverse of that in Part 1, that is, the recognition of the incorrect word or phrase in a sentence. The format of these items, with choices spread throughout the sentence, would appear to require the processing of the entire sentence more than does the format of the first fifteen items, in which the task is focused on a single word or phrase.

Competencies Required:

These items require the full range of grammatical competencies, including lexis, morphology, and syntax. Item 25, for example, requires knowledge of the distinction between "like" and "alike," while item 33 requires the knowledge that the verb "organize" has a lexical selection feature that requires a plural object. An example of an item that requires morphological competence is item 31, which requires the knowledge that the verb form is "produce," and "product." An item that taps syntactic competence is item 19, which requires knowledge of correct word order.

In addition, two items require competence in cohesive reference. The key, "them," in item 30 is incorrect because it refers back to a singular subject, while the key, "the," in item 38, although syntactically correct, does not provide the same cohesive tie that "its" would.

Next, consider item 32, which would appear to require illocutionary competence.

The formation of snow must be occurring slowly, in calm air, and at a
A B

temperature near the freezing point.
C D

Identification of "A" as being in the incorrect tense requires the knowledge that the illocutionary force of the sentence is to define a process. If the illocutionary force is interpreted as that of describing an event, however, the sentence is correct as it stands.

Finally, consider two items that may require sociolinguistic competence. In item 20, the noun phrase "June bugs they" is grammatically acceptable in some dialects of English, although it is unacceptable in the register of the item. In item 28, some purists might find "according to" instead of "depending upon" unacceptable in this register.

Characteristics of context:

As indicated above, the format of these items renders the context much more an integral part of the task required. Irrelevant and unduly demanding context does not appear to be a problem here.

This section of the TOEFL is perhaps the most highly restricted in terms of performance factors, and the most highly focused in terms of competencies required. The most serious problem for this section is the effect of context on the tasks in the first fifteen items. Specifically, although the contexts of some items are reduced and compact and, therefore, probably difficult to process, in most cases the processing of all the information is not critical to successful completion of the task. To the extent that test takers do attempt to process all the information, and to the extent that this requires competencies other than grammatical competence, this will be a source of error in measuring grammatical competence. On the other hand, it might be regarded as a positive characteristic that these items require competencies in addition to grammatical competence. The problem here, however, is that it becomes difficult to identify the exact effect of these other competencies on the students' test performance. One solution to this apparent dilemma might be to provide an extensive and relevant context for all such items, as might be possible with a cloze passage.

Section 3, Reading Comprehension and Vocabulary

Items 1-30, Vocabulary

Restrictions:

The language performance on these items is restricted entirely to single declarative sentences. There is a variety of propositional content, including topics from the sciences, history, economics, government, and literature. The channel is visual.

Task:

The basic task in this part is to recognize lexical synonymy between an underlined word in a statement and one of four isolated words. In all the items except one, the correct match depends entirely upon lexical signification, as opposed to the meaning a word may acquire in a given context.

Competencies required:

Virtually the only competence tapped in this part is lexical competence. One item (20), however, appears to require knowledge of figurative meaning.

> During their heyday, showboats were popular and generally prosperous.
>
> (A) golden age
> (B) infancy
> (C) summer voyages
> (D) revivals

Characteristics of context:

As noted for the first fifteen items in Part 2 above, this section would appear to suffer from misguided attempts at contextualization. Not only is the contextual information in virtually all the items irrelevant to the performance task required, it may present the same problems associated with processing that were discussed above. Consider, for example, item 1, in which the information is both compact and abstract.

> In masculine rhyme, the end sounds of stressed syllables are repeated.
>
> (A) dominant
> (B) vowel
> (C) hard
> (D) final

Here, several abstract concepts, "masculine," "rhyme," "stress," and "syllable," which may be totally unfamilar to the test taker, are introduced in a relatively short sentence. In items such as this (see also items 5, 6, 8, 14, 16, 24, and 29), a successful test-taking strategy might be to ignore the context completely, and rely solely on one's vocabulary knowledge. The items in this section, then, would appear to actually discourage communicative performance, in that the processing of the information in the context may be counterproductive to successful completion of the required tasks.

Items 31-50, Reading Comprehension

This part of the TOEFL has, in my opinion, the greatest potential for requiring communicative language performance. It is less highly restricted than any other part with respect to organization of discourse, propositional content, illocutionary force, and forms. And although the participants are restricted to an "academic" author and a reader and the channel is restricted to the visual, these restrictions are virtually endemic to reading in an academic situation and, therefore, pose no problem of artificiality in themselves.

Tasks:

There are basically two tasks in this part: (1) comprehending a written text and (2) providing requested information based on the content of that text. The five texts in this form of the TOEFL include a range of topics, vocabulary, syntactic structures, cohesive devices, discourse organizations, and illocutionary acts. The questions are of two types: incomplete statements and direct information questions. The type of information requested is both literal and inferential.

Competencies required:

Many of the items in this part appear to require only grammatical competence. Item 33, for example, requires equating the signification of "breathe" with that of "get air," and the recognition that "larvae" is the subject of the verb "getting air" in the appositive clause "getting air through tubes ...." To successfully answer item 36, lexical and syntactic competence are required to equate "came from New York" with "native New Yorker."

The majority of the items, however, appear to tap cohesive competence as well. Item 49, for example, requires lexical competence to equate "lake" with "open water," both lexical and syntactic competence to equate "likely setting" with "usually spend most of their lives," and cohesive competence to equate "allelomimetic behavior" with "it." A similar example is item 59, which requires lexical competence to equate "subtle and frustrating" with "difficult," lexical and syntactic competence to equate "formulation of good research" with "than is generally believed," and cohesive competence to equate "student researcher" with "those who have not actually attempted it," and "a research project" with "it."

One item appears to require illocutionary competence in addition to the other competencies mentioned above. To correctly answer item 47, the test taker must recognize that the illocutionary force of the passage is to define and describe.

Several items appear to require strategic competence, to the extent that this is involved in inference and drawing on relevant extratextual knowledge. In item 52, for example, the test taker must know that cattle are large-hoofed mammals and that bears are not pack-hunting carnivores. In addition, he or she must apply the concept of "mutual stimulation and coordination" to rule out "horses running at a race track" and the concept of "doing the same thing" to rule out "dogs working with police officers."

Characteristics of context:

As indicated above, in this part of the TOEFL, the context is the richest and most relevant to the performance of the required tasks. This context permits a wide range of communicative performance tasks and has the potential for requiring the full range of competencies associated with communicative competence. In addition, the range of topics covered would appear to be relevant to academic settings.

Summary

Restrictions. One question that might be considered in discussing the restrictions on language performance observed in this form of the TOEFL is, "To what extent are these restrictions artificial?" It would seem obvious that those parts of the TOEFL that have no discourse are quite artificial. Lists of isolated, totally unrelated statements are perhaps characteristic only of tests such as the TOEFL.

With regard to the propositional content, most of the topics are of a general or academic nature, which would seem entirely appropriate to the purposes of the test. Of course, one can always argue about the relative interest of a given item or passage, but this is a matter of concern for sensitive item writers rather than for the table of specifications of the test itself.

The majority of the items perform ideational functions, such as describing, stating, defining, and expressing. Many of the items also perform manipulative functions, such as ordering, requesting, and suggesting. As with propositional content, although the illocutionary force of the tasks may be relatively restricted, it may well be that these are the major functions used by students in an academic setting.

There are whole sections of the test that are restricted to a single form, a declarative statement, and it would seem that this restriction results in highly artificial tasks.

With regard to participants, the restriction to author and reader seems reasonable, in that this is the norm for reading. Those parts of the test involving other participants (Section 1, Parts B and C), however, appear quite artificial, largely because the relationship-forming functions necessary to identify these participants are missing.

The restriction to the receptive mode is not in itself highly artificial, given the proportion of receptive language performance that typifies academic study. It does, nevertheless, severely limit the range of performance tasks that can be included in the TOEFL. The Test of Spoken English provides a useful complement to the TOEFL in this respect. The performance tasks on the TOEFL employ both the audio and visual channels, so that a restriction in type of input would not appear to be a source of artificiality.

Finally, the time and space limits placed on the performance tasks, although adequate for the majority of test takers, nevertheless appear artificial, not so much in amount as in principle. One can think of many situations in which there is even less time or space for language performance than on the TOEFL tasks. Rapid conversations in plane terminals, squeezing messages into international cables, and preparing papers for conferences come immediately to mind. At the same time, the idea of having to complete a certain number of items in a given time, or of having to express one's response to an entire discourse with a dark mark on a piece of paper, give that performance an air of artificiality.

In summary, we might ask to what extent the artificial restrictions on the TOEFL affect the test's validity. Obviously, they have little effect on the appearance of validity (face), for the TOEFL, by its very pervasiveness, has become a standard by which other tests are judged. Ironically, the very artificiality that characterizes much of the language performance on the TOEFL has become accepted as the kind of performance that makes a test a test. I believe this is one reason why so many persons react negatively to the cloze procedure--it does not impose the same restrictions that people expect of a test. However, I believe these restrictions do pose a serious problem for the content (domain) validity of the TOEFL and for its construct validity as well.

A second question that might be asked with regard to these restrictions is, "To what extent are they necessary?" First, let us consider the restrictions imposed by psychometric theory. One assumption of test theory, both classical true-score and latent-trait models, is that test items are locally independent. This means that the probability of an individual's getting an item correct is a function only of his or her ability level and the difficulty level of that single item. For this assumption to be met, test developers must write and arrange items so that they are as independent of each other as possible in terms of the tasks required and content included. Clearly, this is at odds with communicative language performance, in which the "items" of discourse are by definition related to each other and to a given context. Two parts of the TOEFL (Section 1, Part C, and the reading comprehension part of Section 3) would appear to involve a certain degree of item dependency, in that groups of items are based on the same text. The questions in these sections, however, appear to be written so as to minimize any overlap in content. And it is largely for this reason that these parts of the test do not realize their full potential for eliciting communicative performance.

A second assumption of currently available latent-trait models is that the test items constitute a unidimensional scale, that is, that they all measure a single trait or ability. This assumption would also appear untenable, not only in terms of current frameworks of language proficiency, but also in light of recent research in testing language proficiency. Indeed, two recent studies (Swinton & Powers, 1980; Dunbar, 1982) strongly suggest that the items in the TOEFL are not unidimensional. As with the assumption of local independence, attempts by test developers to satisfy the assumption of unidimensionality may well result in items which are artificially restricted in their form and content. In fact, the quintessential "discrete-point" item might be regarded as unidimensional. This artificiality would appear to be especially characteristic of the parts of the TOEFL dealing with structure and vocabulary.

Next, let us consider administrative constraints on the TOEFL. There are obvious practical limits with regard to time for administering the test, as well as practical considerations in scoring hundreds of thousands of tests in time for the results to be used by college admissions officers. There are also considerations that derive from the necessity of giving the TOEFL several times a year under reasonably secure conditions. These considerations have to do with the development and equating of multiple forms and are closely related to the psychometric considerations discussed above. In general, these administrative considerations appear to impose restrictions on the number of items and restrict the form to the four-choice multiple-choice format.

But are these restrictions necessary? First, if current theoretical frameworks and research describe communicative language proficiency as involving several distinct but related traits and communicative language performance as occurring in the context of discourse, with interrelated illocutionary acts expressed in a variety of forms, it would seem that language tests would provide both a challenge and an opportunity for psychometricians to test the assumptions of current models and to develop more powerful models if needed. Further, it would seem that there exists a similar challenge and opportunity for test developers to find more creative test procedures and formats.

An excellent candidate would be the cloze procedure. I have recently completed a study comparing deletion types (fixed-ratio and rational) in cloze passages, which has convinced me, at least, that the difficulty of closure is primarily a function of amount of context, in terms of syntax and cohesion not simply number of words, and that difficulty level can be controlled through the use of rational rather than systematic random deletions. This would explain, in part, the inconsistent results that have been obtained from cloze studies using fixed-ratio deletions. These results are also consistent with those of an earlier study (Bachman, 1982) and suggest that the competencies assessed by the cloze can also be controlled, to a large extent, through rational deletions. I am beginning to explore the possibility of using a machine-scorable answer sheet in which each item has twenty-six options, one for each letter of the alphabet. So far I have found that it is nearly always possible to identify one letter within the first six letters of a given set words that will uniquely specify the correct word or words and rule out any incorrect words. This unique letter might be specified, for example, by an asterisk "*" in a sequence of hyphens, instead of a blank, so that "- - - * - -" would indicate that the test taker is to mark on his answer sheet the fourth letter in the word that correctly completes the closure. The advantage of this procedure, it seems, is that it is one step closer to production, in that it requires the recall and not simply the recognition of the correct word.

Another procedure that might be considered is dictation. Although we still score dictation protocols by hand, we have made considerable progress in making the scoring less time-consuming, and it is not unreasonable to expect that recent advances in microcomputers, along with their increasing availability, may provide the means for making this valuable testing technique feasible for large-scale testing.

CONCLUSION

In this paper I have presented a framework for examining performance tasks on language tests and have attempted to demonstrate how this framework might provide some insight into the types of language performance elicited by the TOEFL and into the language competencies it measures. The results of this examination suggest that the TOEFL in general does not require communicative language performance and that the majority of the tasks measure only grammatical competence, while a smaller number tap cohesive competence, with only a handful tapping either illocutionary or sociolinguistic competence.

I have suggested that these characteristics are due largely to the number and types of artificial restrictions imposed on the TOEFL, but the extent to which this is so should be examined empirically. Specifically, it would seem useful to use this framework, along with that developed by Duran et al. (1985) to examine other forms of the TOEFL to determine whether this particular form is representative of the TOEFL in general. Then, to better understand the characteristics of academic language use, these frameworks could be used to examine representative academic discourse, including both written and spoken texts. This could initially be based on nonreciprocal situations, such as lectures and written texts, but might eventually include ethnographic studies of reciprocal language-use situations such as those in classrooms and student-teacher conferences. Such research could also be of use in further specifying referential content areas.

My examination of the TOEFL, as well as that of Duran et al., suggests, in general, that the thrust of any revision of the TOEFL should be toward more appropriate, relevant contextualization of items and more authentic texts. Operationally, a minimum revision might require only minor modifications in the table of specifications and a reorientation of item writers. A full commitment to content and construct validity, however, will probably require elimination of certain parts and item types and the development of other test types. Such revisions will have implications for both test administration and test theory. They will require creative applications of current models and technology and may stimulate the creation of new models and new technology.

As a language test developer and a researcher in language use, I recognize the potential of the TOEFL for providing not only a standard for the measurement of language proficiency but also a tool with which to better understand the nature of that proficiency. I view the convening of this conference as an indication of the TOEFL Policy Council's commitment to assuring that the TOEFL reflects "the most current trends and methodologies in the field," and hope that the suggestions made in our papers and in our discussions at the conference will prove useful in the attainment of this objective.

References

Bachman, L. F. (1982). The trait structure of cloze test scores. TESOL Quarterly, 16, 61-70.

Bachman, L. F., & Palmer, A. S. (1984). Some comments on the terminology of language testing. In C. Rivera (Ed.), Communicative competence approaches to language proficiency assessment: Research and applications (pp. 34-43). Clevedon, England: Multilingual Matters.

Bachman, L. F., & Palmer, A. S. (in press). Fundamental considerations in the measurement of language abilities. Reading, MA: Addison-Wesley.

Dunbar, S. B. (1982). Construct validity and the internal structure of a foreign language test for several native language groups. Paper presented at the annual meeting of the American Educational Research Association, New York.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. E. (1985). The TOEFL from a communicative viewpoint on language proficiency: A working paper (TOEFL Research Report 17). Princeton, NJ: Educational Testing Service.

Swinton, S. S., & Powers, D. E. (1980). Factor analysis of the Test of English as a Foreign Language for several language groups (TOEFL Research Report 6). Princeton, NJ: Educational Testing Service.

Upshur, J. A. (1971). Productive communicative testing: Progress report. In G. Perren & J. L. M. Trim, (Eds.), Applications of linguistics (pp. 435-442). Cambridge, England: Cambridge University Press.

COMMENTS ON THE BACHMAN PAPER

Richard P. Duran

Lyle's analysis of features that influence performance on a test is very useful, although I have some difficulty with his definitions. He tries to distinguish those aspects of performance on a test that are clearly related to the method of testing and the use of particular modalities of language that would affect performance, and he also is interested in particular individual characteristics that would affect performance. I think these are very significant matters, but I am not sure that the things you call methods factors are hard to define. Certainly, the notion of reciprocal and nonreciprocal stands out. What is the character of an interaction with a text? Is there an internal interaction between an individual and a text, or are others involved in the negotiation of meaning in a setting? It is clear that tests such as TOEFL cannot provide a reciprocal in-situation kind of feedback.

A key matter brought up by Lyle is related to some of the findings that we came to in our study of the TOEFL (Duran et al., 1985). We came to the conclusion that sections of the TOEFL that involve more language test a greater range of competencies. Lyle has appropriately pointed out that that conclusion needs to be qualified since it means that an examinee is processing information at a level that is more involved and for a specific purpose. So the conclusion that an increasing amount of language will lead to the assessment of an increasing array of skills is too simplistic. You need to look at the factors that Lyle has cited as they relate to the kind of interaction and the task.

Some of the factors that Lyle has cited are hard to operationalize in terms of any kind of measurement. One example is density of information. Given that the purpose of interacting with the text affects the perception of what information is to be gained, measuring the density of information is a very complicated task because there is more than one level of information in the text.

Another factor he cited is type of information, abstract or concrete. It is very difficult to give these terms dimensions by which they may be defined. Certainly these factors become easier to deal with when we talk about language in a specific context, especially if things that are talked about are things in the environment.

I am concerned about the usages of the term "pragmatics" that have occurred here thus far. My understanding of pragmatics is that it refers to being able to understand the meaning of language in terms of some real-world context. Some of the usages of this term here drift into a cognitive and inference analysis that may not necessarily involve a real-world situation.

In his paper Lyle talks about strategic competence being activiated receptively when a person reads a text and tries to make inferences about its meaning. This notion of strategic competence may be eschewed. As I read the literature on strategic competence, it involves an active capacity to monitor or embellish an interaction to increase the delivery of information to someone else. To view it receptively is a kind of peculiar notion because the kind of feedback that is involved is very different from the kind that is referred to in the literature. We didn't touch strategic competence in our report because we took the position that it involved language production, either in the form of being able to recognize it in others' production, or actually demonstrating it through performance.

Lyle calls for more research on the demands of test items. I think we should do further analyses of the structure and content of items in terms of the factors Lyle talked about and that we talked about in our report (Duran et al., 1985). In general, some analysis of the factors that affect performance should go along with the development of items and with their specifications. Most of the presenters at this meeting have mentioned the desirability of increasing the thematic content of items by linking them or by increasing the contextual information surrounding items. If this were to be done it would be critical to set up a research study in which performance on these new items was compared with performance on the traditional ones. TOEFL general research strategy is to get some information on academic language demands (Bridgeman & Carlson, 1983; Powers, 1986) and then to conduct a validation study wherein the existing TOEFL and new item types are examined in relation to academic language demands. This is a convergent-discriminant validity design, but I think its scope could be increased by increasing the number of skills that are tested and then assessing the gain in prediction that is realized by the new item types. We should also remain open to the possibility that attempts to increase the face and content validity of items may not produce any empirical gain. Thus, attempts to do this may not lead to improvements in terms of construct validity or predictive validity.

One of the major contributions of a communicative approach may be in the tailoring of tests to specific examinees. Thus, in the future we may be able to provide admissions officers with domain-related assessments. However, in addition to admissions purposes, communicative competence tests may help us to verify theories of communicative skills.

References

Bridgeman, B., & Carlson, S. (1983). A survey of academic writing tasks required of graduate and undergraduate foreign students (TOEFL Research Report 15). 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 17). Princeton, NJ: Educational Testing Service.

Powers, D. E. (1985). A survey of academic demands related to listening skills (TOEFL Research Report 20). Princeton, NJ: Educational Testing Service.

A RESPONSE TO LYLE BACHMAN'S "THE TEST OF ENGLISH AS A FOREIGN LANGUAGE AS A MEASURE OF COMMUNICATIVE COMPETENCE"

Charles W. Stansfield

I would like to thank Professor Bachman for a fine paper. I particularly appreciate the way he has interwoven content and construct validity issues with advances and concerns emanating from the field of applied psychological measurement. In my reaction, I will offer a few remarks about the steps the TOEFL program has already taken to make the test more reflective of communicative competence and a few observations about the proper role of context in a test of academic language proficiency. Since Professors Hinofotis, Duran, and Larsen-Freeman also cover these issues, I will focus the majority of my observations on many of the practical suggestions that Professor Bachman makes in his paper.

Professor Bachman finds that the chief fault of the TOEFL is the lack of context associated with many of the items. He is most positive in his evaluation of the reading comprehension portion of the test, and most negative in his evaluation of the statements found in Part A of Section 1. In general, his evaluation of the various item formats in the test agrees with the findings of a study I participated in recently (Duran et al., 1985). Last May an earlier draft of this report was distributed to TOEFL program and test development staff. Subsequently, a meeting was held between both groups at which the findings were discussed. At this meeting agreement was reached to modify the language used in the TOEFL in a number of ways so as to make it more authentic. The changes, which have already been implemented, are as follows:

1. Effective immediately, the test development area will incorporate into reading comprehension pretest questions a variety of alternatives for the standard "according to the passage" leads now in the test.

2. A reasonable proportion of dialogue items will be written in which the questions do not depend strictly on the last utterances. To give greater communicative context, more three-turn dialogues will be created.

3. Sociolinguistic aspects of language, for example, sarcasm, demand, and frustration, can be included if they involve a neutral third party.

4. Extended dialogue exchanges in Part C of Section 1 will contain appropriate natural language leads and insertions so as to avoid the appearance of being expository prose presented orally.

5. In Sections 2 and 3, efforts will be made to include tenses other than present, simple past, and present perfect. Similarly, we will consider reducing the number of passives in each test.

6. Vocabulary items will involve questions as well as statements. The same will apply for written expression items.

7. Among vocabulary items, the number of sentences dealing with American history and geography will be reduced.

8. A greater variety of voices will be used in the Listening Comprehension section of the test.

The Committee of Examiners is currently reexamining the specifications for the Listening Comprehension section. It will be meeting to discuss these immediately following this conference. While I don't wish to speak for the committee members, I suspect they will have a number of ideas for revising the specifications based on the observations offered in Duran et al., the recommendations made at this conference, and their own experience. I mention this because it is relevant to remember that the committee continuously and systematically reexamines the test specifications based on recent input. The changes already implemented are an example of this. The revisions mentioned above, and others that will be made by the committee, will help assure that the TOEFL is reflective of current models of language and language behavior, at least to the degree allowed by available technological constraints.

The role of the degree of context reduction in the Listening Comprehension section is intriguing. Certainly, context is important to face validity and to content validity, but I am not sure how important it is to construct validity. If it is essential to construct validity, we should see this reflected in test statistics. Believing this to be the case, I perused two test analysis reports (Hicks, 1980; Hicks & Morgan, 1980). Test analysis reports contain standard test statistics on each form of the TOEFL, including part-total correlations. I found that for Section 1, Part A correlates highest with section total (about .92), while Part B correlates about .89, and Part C correlates about .86. Part A contains twenty items, while Parts B and C contain fifteen items each. Since our standard test analysis report does not include correlations that have been corrected for spuriousness, the higher correlation for Part A may be due to its greater voice in the section score. However, the fact that Part B shows a higher correlation than Part C with section total does not lend support to the belief that Part C is more valid because it contains a greater amount of context. Although these differences in correlation are small, they are probably significant since they are based on samples of 1,500 examinees or more. It may also be noteworthy that the patterns of part-section correlation are consistent for both domestic and foreign examinees.

This suggests that more understanding of the role of context in examinee performance is needed. One approach for us here at ETS would be to systematically examine test and item statistics to see if degree of context is in any way related to performance. It may be that the variable affects only students with certain background characteristics or students at a certain ability level. Another approach would be to conduct controlled experiments in which examinees would be given items of the same format that include increasing amounts of context. Still another approach would be to determine the efficiency of different item types not currently employed in the TOEFL and to relate their efficiency to their degree of context.

It has been noted that degree of context may affect the authenticity of language and examinee interest in the task. While I agree with this, I am not sure that it affects performance. If it does affect performance, the effect on validity would not necessarily be negative. One could think of many situations in which university students are exposed to language that is not intrinsically interesting, yet the students must still understand the messages being conveyed. Indeed, in the past fifteen years a great deal of research has been done on reduced redundancy tests, which by definition also lack context (Spolsky, 1972, 1973). This research has shown them to be quite valid measures of language proficiency. Such measures also appear artificial in that they are intentionally mutilated in one way or another. This again suggests that more understanding of the effect of context on the validity of a language test is needed. Still, because context is a characteristic of natural language use, I agree that it is a desirable characteristic for a language test. Thus, I find it difficult to disagree with the concerns about language raised in Professor Bachman's paper, especially since he was asked to evaluate the degree to which the TOEFL measures communicative competence.

Professor Bachman also brings up the psychometric assumptions of local independence and unidimensionality. As he has pointed out, these important assumptions also represent restrictions on the alternatives available to a responsible test maker. Test scores are derived by equating performance on the current form of the test with statistics from previous forms. Through the application of appropriate equating procedures, performance on forms that vary somewhat in difficulty can be linked in a manner that will be supported by specialists in mathematical statistics. However, if the test format or items violate the assumptions on which the statistical procedures are based, the application of the procedures would be open to debate. Similarly, if the test data itself do not meet the assumptions ascribed to the model, the meaning of scores is open to question.

The concept of local independence of items is important for many practical considerations. It relates to content sampling, since independent items sample the content better than items that are linked, and to the estimate of a test's reliability, since this depends in part on the number of items in the test. As Professor Bachman points out, this is clearly at odds with communicative language performance, where the demand for interaction with the context requires that the items be linked to some degree. This poses no small dilemma for makers of multiple-form tests. While we must continue to work to resolve it, we may have to live with it for a number of years. The introduction of new item types must be accompanied by supporting psychometric theory, and new theories and their validation do not come easily or quickly. In terms of TOEFL, we have spent the past seven years trying to reach a more precise understanding of the implications of the use of item response theory with the TOEFL population and the current format of the test.

On a more optimistic note, it does appear possible to construct locally independent items that are based on a common context. At present, we do this quite adequately with the minitalks and with the reading comprehension passages. Here items are written so as to test independent information. As a result, performance on these individual items is only very slightly more related to performance on other items within the same passage than it is related to performance on items in other passages. If item writers can do this with these contexts, then perhaps they can write items that are locally independent with other contexts, such as the cloze. One approach would be to introduce only one deletion per sentence. This would guard against introducing dependence, although it would not ensure independence. Another approach would be to write items whose responses depend on interpretation of the text as a whole and not on any previous response. Indeed, it may be more feasible to develop skill at writing such items than to await the development of new psychometric theory.

Meeting the unidimensionality assumption poses another challenge for a large testing program such as TOEFL. The assumption considers that the test population is routinely homogeneous and that the test measures a single underlying trait, construct, or factor. The problem of a single construct is resolved by separately equating data for the three sections of TOEFL. Thus, we are able to measure three different constructs and produce a total score based on the total of the three section (construct) scores. On the other hand, our test population is in no way homogenous. Rather, it is composed of at least 160 native language groups whose proportionate representation varies from administration to administration.

Performance on the TOEFL, both on the total test (Swinton & Powers, 1980) and on individual items (Alderman & Holland, 1981), is related to native language group membership. During 1982 and 1983, the TOEFL Research Committee, composed of language testing researchers affiliated with different institutions, devoted several meetings to a discussion of this and reached the conclusion that a modest relationship between native language and performance on individual items should not be a cause for concern. Among advanced learners of English as a second language (with FSI/ILR level 3+ skills, for example) one would expect to encounter different weaknesses in pronunciation, lexis, and syntax, depending on the learners' native languages, even though they were basically equal in overall proficiency. Since a learner's native language plays a role in performance in many language tasks, this will surely be reflected in statistics on individual items, even after an adjustment has been made to allow for differences in overall language proficiency that may exist between native language groups.

The relationship between native language group and performance on the total test is a more complex problem. Differences in total score reflect differences in knowledge of English between among language groups. However, factor analyses raise a different concern. Swinton and Powers (1980) found that the factor structure of Sections 2 and 3 of TOEFL varies by native language group. This means that for different groups, these section scores measure different components of proficiency. For example, for most non-Indo-European language groups, Reading Comprehension items relate more closely to Structure and Written Expression items, while Vocabulary items form a separate factor. This difference in the factor structure of TOEFL can affect judgements concerning its construct validity. The problem is of concern to the Research Committee. It was the principal motivation for their funding a study of the cloze procedure (Hale, 1983) designed to determine whether multiple-choice cloze items (or a subset of them) relate more consistently to reading comprehension than do vocabulary items. If it turns out that they do, then the replacement of vocabulary items with cloze items would produce a more uniformly unidimensional score on Section 3.

I mention these points because I do not believe that the assumptions of unidimensionality and local independence can be taken lightly. Like Professor Bachman, I believe that language tests provide a challenge for psychometricians to develop more powerful models for scoring and equating tests (that is, ones that can be used with a heterogenous population) and for test developers to create valid item formats that will conform to the models. ETS psychometricians, as well as others around the world, are now working on the development of multitrait models. Similarly, the TOEFL program is experimenting with new item types such as the multiple-choice cloze. We hope these efforts will be fruitful.

Since I have twice made reference to our experiment with a multiple-choice cloze, let me tell you a little more about it. Next month, some 25,000 examinees taking the TOEFL in North America will receive a four-section test. The fourth section will consist of a fifty-item cloze test constructed from three passages that appeared previously in the reading comprehension section of the TOEFL. The passages were selected by Professors Oller and Hinofotis, who also selected appropriate deletions using a rational deletion procedure. They then constructed suitable distractors so that each item offers four options. Some of the items and options that were selected emphasized sensitivity to sentence level constraints, while others emphasized discourse level constraints. Similarly, the above categories of items were further subdivided to identify those that focused principally on either vocabulary or syntax. After the passages were field-tested with students at UCLA, Hunter College (CUNY), Rutgers, and Columbia, Gordon Hale and I made revisions in them based on the preliminary item analysis and suggestions from TOEFL test development staff.

The experimental cloze passages will be administered to all TOEFL examinees in the United States and Canada at the November 1984 test administration. Through the use of factor analysis, the data from this administration will be used to determine the role of cloze items in the factor structure of the TOEFL. The data will also be subdivided by native language groups. Another major aspect of the statistical analysis will be an attempt to determine whether item response theory parameter estimation methods can be applied to multiple-choice cloze items. If the probability of a correct response, as specified by the item response functions, corresponds to observed examinee performance, this would provide support for the potential use of IRT equating in an operational form of the TOEFL that might include multiple-choice cloze items.

I am pleased that Professor Bachman sees value in the use of a rational deletion cloze. It makes me feel more confident that we are on the right track. However, I note that the cloze he describes would involve exact word scoring. This type of scoring procedure is fine for research, since it is reliable and convenient (that is, it can be machine scored). However I don't believe that it should be used with tests that have a bearing on important decisions about examinees, since there may be more than one acceptable answer.

Another concern I have with the type of cloze procedure Professor Bachman describes, at least upon initially learning about it, is that each item has twenty-six options. I would imagine that this would make it rather easy for examinees to miskey an answer. A more foolproof technique would be to have a scanner recognize a letter printed by the examinee; however, the development of pattern recognition technology is still in its infancy.

The use of dictation via microcomputer is another idea suggested. The context provided by dictation is the principal reason why attitudes toward its validity have changed so radically over the past two decades (Stansfield, 1985). In the early sixties, Lado (1961) wrote that dictation measured "very little of language" because "the words in many cases may be identified by context if the student does not hear the sound correctly" (p. 34). More recently Cohen (1980) devoted eight pages of his textbook on language testing to a discussion of the value of dictation. According to Cohen, dictation's dependence on contextual clues makes it a valid measure of functional language ability. As suggested by Professor Bachman, the student may be able to type a printed version of a dictation on a microcomputer. I should mention, however, that it will be many years before microcomputers are readily available to large numbers of examinees in the developing nations, where most of TOEFL's test centers are located. Research would also be needed on scoring procedures. If partial credit were given to examinees who write in the correct word but misspell it, this would amount to a system of weighted responses for each item, thus introducing this new variable into the calculation of the test score.

The discussion of the possible use of microcomputers raises the topic of other computer applications to the TOEFL program, in particular, computerized adaptive testing (CAT). CAT offers many advantages over conventional testing. It permits the creation on site of an individually tailored test and the calculation of a score that can be immediately presented to the examinee. Another advantage of CAT is the possibility of using new item types, that is, those that are more oriented toward production. Thus, returning to Bachman's idea of a cloze test that is not based on recognition of the correct response, CAT would allow the examinee to type the response on the terminal. The use of weighted responses (according to native speaker norms) might also be possible given adequate research on score equating. It may even be possible for the computer to interact with the examinee through a simulated conversation in which each would have to adjust to the other's input.

The advent of computerized adaptive testing offers many possible advantages to language testing, as well as testing in general. Someday, it may even be possible to simulate a direct oral interview using interactive video-disc technology. This day may be several decades away, however, and for the moment we must focus our attention on the present TOEFL. Professor Bachman's paper is a contribution to our understanding of the present TOEFL and it has implications for changes that can be implemented during the immediate future.

References

Alderman, D. L., & Holland, P. W. (1981). Item performance across native language groups in the Test of English as a Foreign Language (TOEFL Research Report No. 9; ETS Research Report No. 81-16). Princeton, NJ: Educational Testing Service.

Cohen A. D. (1980). Testing language ability in the classroom. Rowley, MA: Newbury House.

Hale, G. A. (1983, October). The role of cloze items in the TOEFL (Proposal submitted to the TOEFL Research Committee.) Princeton, NJ: Educational Testing Service.

Hicks, M. M. (1980, March). Test analysis, TOEFL form 3ATF8 (Statistical Report 80-22.) Princeton, NJ: Educational Testing Service.

Hicks, M. M., & Morgan, R. V. (1980, May). Test analysis, TOEFL form 3CTF5 (Statistical Report 80-74.) Princeton, NJ: Educational Testing Service.

Lado, R. (1961). Language testing. New York: McGraw Hill.

Spolsky, B. (1972). Redundancy as a language testing tool. In G. E. Perren & J. L. M. Trim (Eds.), Applications of linguistics: Selected papers of the Second International Congress of Applied Linguistics, Cambridge, 1969 (pp. 183-190). Cambridge: Cambridge University Press.

Spolsky, B. (1973). What does it mean to know a language or how do you get someone to perform his competence? In J. W. Oller, Jr. & J. C. Richards (Eds.), Focus on the learner: Pragmatic perspectives for the language teacher (pp. 164-176). Rowley, MA: Newbury House.

Stansfield, C. W. (1985). A history of dictation in foreign language teaching and testing. Modern Language Journal, 69(2), 121-128.

Swinton, S. S., & Powers, D. E. (1980). Factor analysis of the Test of English as a Foreign Language for several language groups (TOEFL Research Report No. 6; ETS Research Report, No. 80-32). Princeton, NJ: Educational Testing Service.

DISCUSSION OF THE BACHMAN PAPER AND THE RESPONSES BY DURAN AND STANSFIELD

Lyle Bachman. Let me first respond to some of Richard Duran's comments regarding the definition of performance factors. I certainly can't provide a clear definition of density of information or criteria for distinguishing between concrete and abstract information. However, in the process of reviewing the TOEFL and other tests, I have a much better feel for what these terms mean. So, in the long run, I think we will be able to define them operationally. I view strategic competence as the use of compensatory strategies. But we also see strategic competence at the upper level in the articulate speaker. In terms of the TOEFL, I suggest that whenever a person has to draw on outside information and bring this information to bear on the task of understanding the text, that may be a function of strategic competence as well.

Henry Holtzclaw. One thing that concerns me about computerized testing is the possibility that student scores will be influenced by the ability to type.

Charles Stansfield. I might mention that another study the TOEFL Research Committee has recently funded is one by Marilyn Hicks (1984) involving the development of a computer-delivered TOEFL that will be partly adaptive. In conjuction with this study, we will be asking the examinees who take this test to fill out a questionnaire that will provide us their impressions and some concrete information on what it is like to take a TOEFL via comptuer. One of the questions we will be asking them relates to their experience as typists or at the computer terminal. We can then use this information to begin to address the issue that Henry mentions.

John Haskell. I notice that both Lyle and Charlie seemed to prefer appropriate word scoring of cloze tests. I am curious as to why you view this as more desirable. Also, don't the distractors associated with the multiple-choice cloze make this format less desirable since they are completely inappropriate responses?

Charles Stansfield. Of course, exact word scoring could only be used on the TOEFL if you had a fill-in-the-blank type cloze, which would necessitate computer delivery with the examinee typing in the correct response. A scanner could not read a response that was handwritten by the examinee. While exact word scoring may facilitate hand scoring, on a standardized test it would pose a threat to face validity.

In a multiple-choice cloze, the distractors would not be completely inappropriate. They would be appropriate to varying degrees, either syntactically or in terms of cohesion. They may relate to other concepts brought up in the passage. The multiple-choice cloze represents a gain over the fill-in-the-blank cloze because of our ability to use the context in the construction of the distractors.

John Oller. The point about the independence of cloze items is a critical technical issue. John Carroll argued years ago that he didn't think it was a threat to the cloze. He thought that cloze test items were sufficiently independent to meet the requirements of standard classical theory. At UCLA, Brown (1983) showed that there was no significant difference in reliability of cloze tests when different methods of estimating it are used. And since reliability is related to the independence assumption, he felt that it supports Carroll's observation.

Charles Stansfield. Certainly, on the face of it, one would assume that cloze items are interrelated. However, we have shown cloze tests to psychometricians here at ETS and asked them if they thought the items were independent. Generally, they have felt they are sufficiently independent that they do not violate the local independence assumption. So I am glad to know that after studying the issue, Brown came to the same conclusion.

Lyle Bachman. The problem with the assumptions of local independence and unidimensionality is that they are assumptions. People don't understand the consequences of violating these assumptions. There isn't any empirical way of determining whether items are locally independent. It is a matter of judgment; you can't look at inter-item correlations. Finally, I don't think Brown's research is capable of addressing the issue of local independence.

Diane Larsen-Freeman. Lyle has suggested that we consider adding a cloze and a dictation to the TOEFL. What do you think these tests will contribute to the TOEFL that is not already being measured? Also, do you think these measures would make the TOEFL appear to be measuring communicative competence? It strikes me that they would not.

Lyle Bachman. That is an empirical question. I gather that one of the current concerns with the reading comprehension section of the TOEFL is that it is multidimensional across language groups. I don't see any reason to think that the cloze would perform any differently. So I am not sure that a cloze would provide a solution to that problem. I think that what the cloze might provide would be a form of "window dressing" if the context were authentic. People would see that you are trying to test communicative competence. The importance of other tests such as the dictation and the computer-delivered format lies in the fact that they provide alternatives to the standard multiple-choice format. Although they may not make an empirical contribution to the assessment, at this point you can't go wrong by exploring other procedures. I would also like to say that I don't understand why students initially react so negatively to the cloze. In time, I think they would become accustomed to this format.

Richard Duran. I am wondering if it would be possible to identify the types of activities that a foreign student would be involved in at a university?

Chris Candlin. The British Council's English Language Testing Service (Seaton, 1985) is premised upon creating test modules for particular academic disciplines. I should say that their approach is not unproblematic and that it has not been validated. There were a number of studies conducted at the University of Lancaster in the mid 1970s under the impetus of an ESP project in Jedda, Saudi Arabia, in which the skills necessary for academic reading were addressed. Bill Murphy and I conducted an empirical analysis of lectures given in academic settings. Richard Mead of Birmingham did a report on seminar interaction strategies. All of their data are available, as are extensive quantities of unanalyzed data. Of course, we should remember that there may be differences in academic settings between the UK and the US and that testing was not the motivation for these latter studies.

References

Brown, J. D. (1983). A closer look at cloze: Reliability and validity. In J. W. Oller, Jr. (Ed.), Issues in language testing research (pp. 237-250). Rowley, MA: Newbury House.

Carroll, J. B. (1959). An investigation of cloze items in the measurement of achievement in foreign languages. Cambridge, MA: Harvard University Laboratory for Research in Instruction.

Hicks, M. M. (1984). Development and investigation of computerized and paper-and-pencil placement tests for the TOEFL via two-stage testing procedures (Proposal submitted to the TOEFL Research Committee). Princeton, NJ: Educational Testing Service.

Seaton, I. (1985). Issues in the validation of the English Language Testing Service (ELTS) 1976-1983. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), Second Language Performance Testing (pp. 111-129). Ottawa: University of Ottawa Press.

COMMUNICATION THEORY AND TESTING: WHAT AND HOW[1]

John W. Oller, Jr.

This paper deals with two main questions and their relationship to a third: (1) What is communicative competence? (2) How can it be tested? And (3) What implications do answers to the former questions have for the TOEFL and TSE (but especially for the written portions and the Reading Comprehension and Vocabulary section of the TOEFL)? Related to the first two especially are many mysteries in search of an adequate theory. Where do memory, consciousness, intelligibility, intelligence, authenticity, aptitude, achievement, knowledge, beliefs, attitudes, intentions, primary and nonprimary language proficiency, and the like fit in? How do they relate to communicative competence? Two definitions of communication are considered--(a) the standard dyadic definition and (b) a more general definition in relation to which a pragmatic theory of communicative competence is elaborated. It is claimed that the general definition has certain critical theoretical advantages and that a modular generative hierarchy has advantages over taxonomic approaches. The idea of pragmatic equilibration--dynamic fitting of texts to facts--is proposed (with bidirectional and increasingly effective interaction over time). The theory suggests that the meaningful scalability (i.e., practical validity) of a test (or item) will depend on the degree to which the factual domain (that is, the referential domain including intrapersonal and intergroup relationships) is fixed. Some observations are offered on the present format of the TOEFL. It is concluded that the subtests may be better than we thought as measures of communicative competence and that the means for making them better still are within reach.

It seemed at the start that this paper might be more controversial and more innovative than it has turned out to be. When the outline was devised for the first draft, marked differences between theoretical approaches were anticipated, but they did not show up under careful scrutiny. This is a pleasant surprise and augurs well for the potential of meaningful interaction in a spirit of mutual cooperation and problem solving. The practical solutions to testing problems are converging on certain conclusions that, although yet not constituting a coherent theory, seem increasingly to point toward such a possibility. The taxonomic approaches adopted heuristically by researchers worldwide, perhaps more as a stopgap than as a theory, are constantly enriching the basis for a quantum leap forward to a new level of theoretical coherence. In the interest of inching toward that forward leap, the present paper is offered for the Second TOEFL Invitational Conference.

It might be argued that the taxonomic approaches to the characterization of communicative competence have begun quite naturally from the bottom up. Theorists have examined particular problems and contexts of communication and worked up from there to various preliminary taxonomies of skills, aspects, components, and so on that might lead eventually to a coherent theory. Of course, the theorists who have opted for the bottom-up approach in the tradition of Hymes (1972) are also well aware of the possibility of working as well from the top down. By combining these programmatic options, it should be possible to work both ends against the middle. With this aim in mind, I would like to try in this paper to explore some of the theoretical issues and, in the end, hope to arrive at some concrete suggestions about the present TOEFL format and possible ways in which it might be strengthened as a measure of communicative competence.

Prior Considerations for a Top-Down Theory-Building Approach

As David Lightfoot (1982) notes in his book in search of a biological theory of grammar, different goals usually lead to different research agendas and often to different interpretations of the same empirical findings. Therefore, if prior agreement is attained on what the objectives are, it may be, and probably will be, easier to agree on what sorts of theories and interpretations are relevant to particular empirical findings. Granting that the same research agenda may have different goals when viewed by different interpreters, in the interest of mutual understanding it may be useful to state from the beginning what the goals might be and which ones are likely to be mutually agreed to and which ones are not.

To begin, let me try to state what I see as the mutually accepted goals of this conference. The most obvious goal, perhaps the clearest, is (1) to understand the quality of the TOEFL test with respect to the construct of communicative competence, and beyond this, (2) to improve the TOEFL whenever possible.

These might be termed the primary goals of this conference. However, we all share other interests that are more deeply rooted in biological, psychological, sociological, and educational concerns and that exceed the limits of the TOEFL per se. For instance, there are certain secondary goals that all of us are also interested in to some extent. One of them is (3) to advance understanding of communicative competence as a theoretical construct, and beyond this (4) to achieve a better understanding of how communicative competence is acquired and/or modified by education.

Beyond these aims, our paths may diverge somewhat, but I suspect that most of us share a more general tertiary goal: (5) to understand the distinctive nature of human communication and the role that language plays in it.

And beyond this goal, there may even be common inclinations toward a still more general quaternary goal: (6) to understand the nature of semiotic systems (representational systems in general) and the role(s) they play in communication.

This last objective may call for a bolder research program and theoretical perspective, but it may offer certain rewards in relation to the practical needs of educational programs, including language instruction, testing of communicative skills, and, more specifically, evaluating and continually upgrading the TOEFL.

Although goals 5 and 6 are not everyone's priorities for this conference, bearing them in mind for the top-down approach may help to define the sort of theoretical perspective that will maximize progress in the pursuit of goals 1-4. In fact, it may even be the case that theoretical perspectives of narrower scope may prove less adequate because they are less general.

As Einstein (1954a) observed in 1936 concerning theories of physics, the tendency was "increasing simplicity of the logical basis" (p. 96). The chief advantage sought was internal coherence--in his words, "the internal perfection of the system" (p. 96). Although we are far from achieving the theoretical completeness that a few physicists dared to hope for in the 1930s, there may be benefit in giving thought to a theoretical perspective that aims to embrace the whole scope of communication and communicative competence.

I. Foundational Questions

Two basic questions are distinguished: First, what is communicative competence? Second, how can it be tested? These questions are actually related to a host of others in such a way that they constitute two grand families of interrelated mysteries. These are explored in this section along with two possible definitions of the term "communication." In the next section an attempt is made to sketch out a theory that fits the broader of the two definitions considered.

A. What Is Communicative Competence?

Asking what communicative competence is suggests a more fundamental question--namely, what is communication? This question leads directly into a maze of basic theoretical puzzles. For instance, how is communication related to comprehension, consciousness, understanding, intelligence, comprehensibility (intelligibility), first language proficiency, second language proficiency, foreign language proficiency, knowledge, memory, belief, attitude, intention, and experience? Although the unitary factor hypothesis (a psychometric heresy) has indeed been forcibly retired, it remains true that these constructs are intricately interrelated. Depending upon the definition of one or another, different results may follow for the rest. This is due to the fact that communication is entailed by or itself entails all the rest. To this extent, it seems that an adequate theory of communication will need to show how all these theoretical constructs are interrelated. Therefore, such a model, or theory, might provide one source of hypotheses (working from the top down) about the second family of mysteries.

B. How Can Communicative Competence Be Tested?

Among the corollary questions associated with the basic problem of testing communicative competence is whether or not intelligibility is possible without communication and, conversely, whether or not communication is possible without intelligibility. Similarly, can there be a test of knowledge, or memory, or intelligence, and so forth without the same test also being, to some extent, a test of communicative competence? And conversely, can there be tests of communicative competence that do not to some degree tap many of the other mental constructs that are of concern to educators? For instance, is communication in the normal human sense possible without some sort of intelligence that is generally applicable to a potentially infinite variety of contexts? If so, where? If not, then what is the understood relationship between intelligence and intelligibility?

Depending on the particular theoretical approach chosen in the _definition_ of communicative competence, tests thereof may be defined in different ways. In the following section, two options are explored: the standard dyadic definition of communication and a more general definition.

II. Definitions of Communication

To demonstrate still more explicitly the indissociable characterizations of communicative competence and tests thereof, it will be necessary to be more explicit about the definition of communication per se. Just what is communication? Needless to say, if we go to different theorists, we will get different answers. In fact, the degree of difference among at least some of the theorists was apparent at a recent language assessment symposium convened by the National Institute of Education and Inter-America Research Associates. Rivera (1983) notes that "it was not possible to reach a consensus . . . [on] a working definition of communicative competence" (p. xiii).

A. The Dyadic Definition

However, it may nonetheless be observed that the most common definition of communication assumes _at least a dyad of interacting persons_. This may be referred to as the _standard_ or _dyadic definition_. More explicitly put, it suggests that communication is an interaction between _n_ persons where _n_ must be equal to or greater than two.

1. The Notional/Functional Approach

If I understand the proponents of the notional/functional approach to language instruction correctly, the dyadic definition is the one they generally prefer. For instance, in his discussion of communicative testing, Andrew Harrison (1983) says, "communication . . . necessarily involves expression of information by one part [sic] to the exchange and understanding of it by the other" (p. 78). There are also certain key phrases that suggest a fairly clear definition. Harrison speaks of "the bridging of an information gap" and of "the background information that is used by parties to a real communicative exchange" (p.77). Although not all proponents of notional/functional approaches would agree on the details of Harrison's approach to testing communicative competence, most would probably accept his standard definition of communication as involving at least a dyad of persons engaged in some sort of exchange.

2. The Canale Model

The fact that Canale (1983) seems to have in mind the standard dyadic definition of communication is apparent in his statement that "performance on authentic communication tasks is not always a good predictor of performance on academically oriented autonomous tasks presented in the second language" (p. 338). More explicitly, Canale cites Morrow (1977) and others as arguing that "communication is primarily a form of social interaction in which emphasis is normally placed less on grammatical forms and literal meaning than on participants and their purposes in using language--i.e., on the social meaning of utterances" (p. 340). Canale goes on to say that "authentic communication thus requires continuous evaluation and negotiation of various levels of information" and here he cites Candlin (1981), Haley (1963), and Hymes (1972). Again, it would seem that the standard definition is understood.

3. Other Theorists

It may simply be added here that nearly all other theorists accept some version of the dyadic definition of communication. For instance, Chomsky in all of his writings so far as I know, but especially in 1975 and 1980, consistently advocates the standard definition. He contends that it would be meaningless to suggest that one "communicates with oneself" (personal communication). Jim Cummins too has, according to my understanding of his writings, adopted the more or less standard view, (see especially Cummins, 1983) following Dell Hymes (1972).

B. A General Definition

However, in spite of the popularity of the dyadic definition, a more general definition is tacitly implied by it. This more general definition might be paraphrased in a variety of ways. For instance, we might say that any activity that results in a nonnecessary decrease in entropy (in the sense of information theory), or a nonnecessary increase in orderliness (information) is communication. Or, another way of arriving at a general definition would be to say that communication is any interaction between systems of intelligence, or components thereof, where information is rendered accessible by translation between semiotic (representational) systems. Either of these formulations, I believe, has the effect of equating communicative competence with what is ordinarily termed general intelligence. Incidentially, as I understand them, this is roughly the approach followed by theorists who have taken the routine of pragmatism and pragmaticism as discussed in the writings of Peirce, James, and Dewey. (For references and specific works, see Oller, 1986; also, Watzlawick, Beavin, & Jackson, 1967.)

III. The General Definition and Its Expansion into a Theory

It should be clear that the dyadic definition--which assumes that the interacting intelligences must be human beings and that therefore there must be at least two of them to achieve genuine communication--is a special case of the general definition. The more general formulation would allow that recalling information from memory, storing information in memory, considering alternative courses of action, and in fact all sorts of intrapersonal (or intrasystem) interactions are also communication. However, some would argue that this expansion in the scope of the definition of communication is undesirable because it confuses distinct processes. The important question is whether this undesirable consequence can be avoided, and at the same time whether or not there is anything to be gained from such a broad definition of communication. We will consider these issues by examining what is to be gained first and then returning to the resolution of the potential difficulties in an expanded theory of communicative competence, alias general intelligence.

A. Advantages of a General Definition

The wonder of human communication is the fact that private experiences, understandings, and acts are rendered publicly accessible. Further, it seems that the accessibility of private information to the public at large is limited only by the intelligence (and sincerity) of the communicators and by constraints on the semiotic systems used to achieve communication. If the semiotic systems are completely general and if they are nonfinite, then spatiotemporal constraints would seem to be the only limitations preventing full and complete communication. However, by the same logic, dyadic communication (or communication according to the standard definition) cannot occur at all apart from the possibility of communication in the more general sense of the term. That is to say, dyadic communication is completely dependent upon interaction between semiotic systems such that nonnecessary decreases in entropy (or increases in information) are brought about in intelligent systems. In other words, the dyadic definition presupposes the more general definition and without it is meaningless.

What is more, if we take particular cases of communication of the dyadic sort, we discover that there are private, intrapersonal, aspects from the separate view points of both interlocutors that are just as important to the dyadic, interpersonal interaction as the aspects that are intentionally made public. For instance, consider an incident some years ago involving a deer hunter who heard a police siren and, having nothing better to do, watched through the telescope on his rifle as a state patrol officer stopped a speeder on the highway several hundred yards below the hunter's vantage point on a hillside. The speeder pulled over. The officer followed suit, got out, and approached the speeder's car. Suddenly, what began as a routine event went haywire. The hunter heard the report of a large caliber pistol and saw the officer fall. The assailant stepped out of the vehicle, aimed his pistol at the fallen officer, and proceeded to shoot him again and then again. As soon as the hunter saw that the assailant intended to kill the officer, he had to decide whether or not to take countermeasures. In order to take account of the possibility of countermeasures, it was necessary for the hunter to see (understand) what was happening on the road below and to take account of his own weapon that could potentially equal the odds in favor of the fallen man. Within a heartbeat or two, he released the safety on his rifle, aimed, and fired, killing the assailant. Later he was tried for murder.

Now, suppose we look at this incident from the point of view of the dyadic definition. We would have to say that some communication took place between the police officer and the speeder, but not between them and the hunter until the point where he shot and killed the assailant. Even then, it would be difficult according to the dyadic definition to say that any communication occurred (although, in some sense, an information gap between them was bridged, and both were changed by the interaction).

On the other hand, by the general definition, a great deal of communication occurred prior to the shooting. Not only did the officer communicate (intentionally) with the speeder by pulling him over with the siren, but he also communicated (unintentionally and incidentally) with the hunter, who understood the meaning of the semiotic system and its relation to laws concerning speed limits and the like. Further, not only did the speeder communicate with the officer by shooting him (intentionally), he also (unintentionally and incidentally) revealed to the hunter his plan to kill the officer by stepping out of the car, pointing his gun at the officer, and firing two more shots.

To the extent that all actions (or nonactions) of intelligence(s) are symptomatic of intentions, beliefs, attitudes, feelings, and the like, it may be claimed, as Watzlawick et al. (1967) have argued, that it is impossible <u>not</u> to communicate. However, in spite of the fact that this axiomatic conclusion may be nontrivial, there are other ways in which intrapersonal interactions are critical to interpersonal interactions and without which the latter would be impossible. This other aspect of intrapersonal interaction must be included within the scope of the term "communication" or, it seems to me, the very possibility of communication in the common sense of the term, in the dyadic sense, will evaporate.

Communication often occurs when we do not want it to. Intentions an interlocutor would rather conceal may be revealed in clearly interpretable semiotic forms. As Jose Marti put it, "Cuando quiero llorar, no lloro; y a veces lloro sin querer." Feelings often manifest themselves in spite of the will of the communicator, and to some extent they may be, as John Dewey and others have observed, as surprising to their originator as they are to anyone else. That is to say, the critical elements of the dyadic definition are present within the experience of a single individual. In other words, a dyad of persons is not required. Rather, what is necessary in order for communication to occur is _a dyad of semiotic systems plus an intelligent interpreter who is able to translate between the systems_.

In human beings, of course, the systems are _never_ simply dyadic. For instance, in order for the hunter to infer from the pointing of the gun and its firing that the assailant intends to kill the policeman, there must be a translation from a visual scene into some other sort of general propositional logic. Similarly, for the policeman to switch on his siren and flashing lights requires a translation from visual information to motor commands resulting in auditory and visual signals that, again, when translated, convey the propositional meaning that the officer wants the speeder to pull over. For the speeder to understand that the wailing siren means "pull over" involves a translation from an auditory signal into a general propositional system in which knowledge of the speed limit is expressed, as well as translation ultimately into a system of motor commands that results in the action of pulling off to the side of the road. However, it would be naive to fail to note that the act of pulling over is intended to convey compliance to the officer while the speeder is actually planning to shoot him.

These observations, moreover, are generally applicable to dyadic interactions. We cannot say that we _understand_ the communications (in the dyadic sense) unless we first understand their underpinnnings in terms of intrapersonal communications. Also, if goal 6 above is taken seriously, it follows that much will be lost if we opt for the narrower definition of communication. There would be no way, for instance, to take account of communication systems that operate either above or below the inter-organismic level--for example, communications between nonpersonal societal structures such as the different branches of a government or departments of an institution, between the neurological and endocrine systems of an organism and its musculature and organs, between the cells of a developing embryo that will become distinct organs, between the proteins and metabolic processes, transport systems, and the like, or between the DNA of a cell and its proteins.

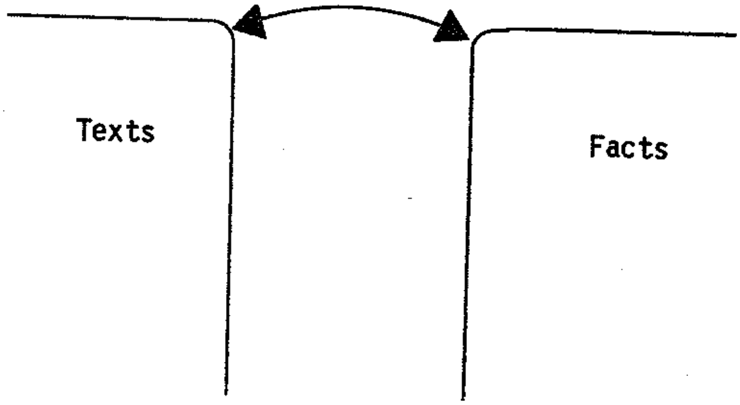
B. Removing the Potential Disadvantage of a General Definition

Having shown that certain critical advantages are gained, it remains to be shown that the potential disadvantage of a general definition, especially the danger of confusing distinct concepts and processes, can be avoided. To demonstrate this will require some elaboration of the theoretical framework. In particular, it is necessary to show how certain critical constructs such as memory, knowledge, belief, sensation, consciousness, semiotic representations, texts, facts, language proficiency, grammar, and so forth can be both distinguished within and integrated into the theoretical model. Beyond this, it will be necessary to show how intrapersonal communication relates to interpersonal communication, though they remain distinct from each other.

Figure 1 gives a rough sketch of the pragmatic interaction that forms the basis for the theory to be elaborated. It is argued that the principal activity of intelligence can be construed as the equilibration or fitting of texts (in some semiotic system) to facts (in some other semiotic system). Roughly interpreted, we may think in terms of the texts of a particular language as related to particular constellations of facts that may be thought of as patterns of events arranged in succession in the stream of experience). According to this theoretical perspective, Figure 1 may be thought of as a summarial representation of what general intelligence does nearly all of the time. It attempts to interpret (make sense of) the facts of experience by relating them to (translating them into) various sorts of semiotic representations (roughly subsumed in Figure 1 under the term "texts").

Figure 1

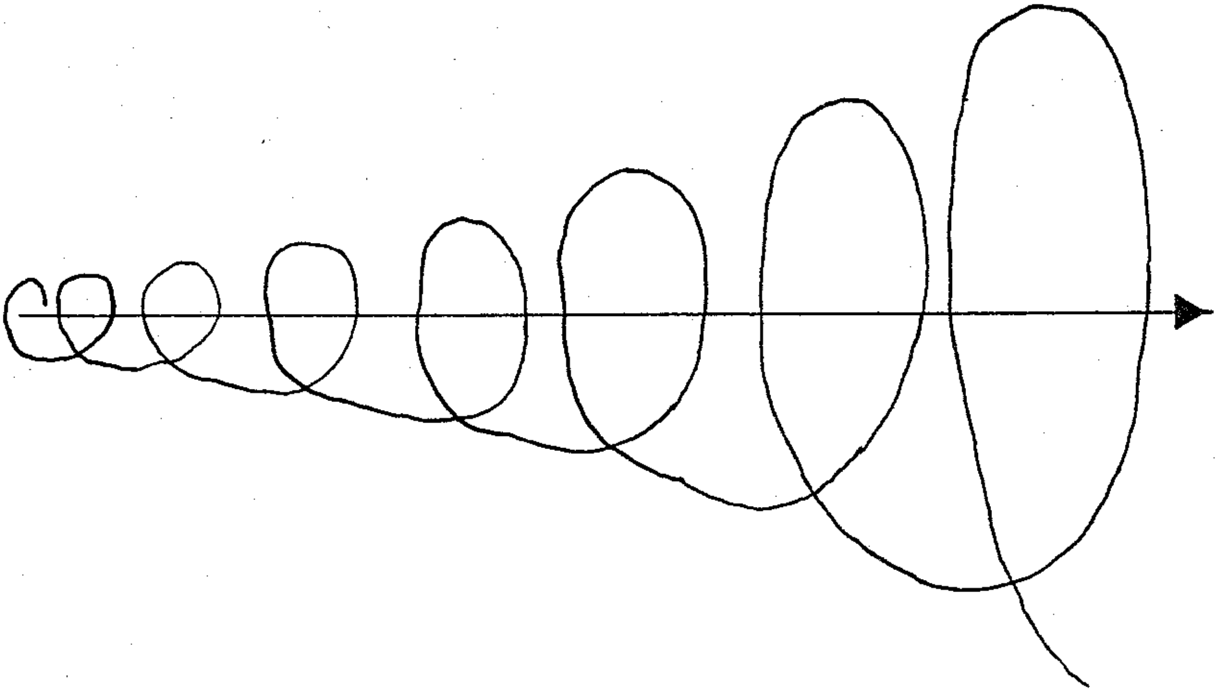
The process of pragmatic mapping in its simplest conceptual form.



Of course, the "facts" in Figure 1 are themselves dependent on a deeper semiotic system that itself imposes textual characteristics on experience. Therefore, the operation of intelligence in pragmatically linking texts to facts and facts to texts is really a process of communication in the sense of the general definition proposed above. It is a process of translation between distinct semiotic systems. Also, it may be observed that the efficacy of this translation typically improves over time. Whether we think in terms of the acquisition of the primary language, or in terms of the development of intelligence in general over the course of growth and maturation, the scope of understanding in normal humans develops something along the lines of the spiral shown in Figure 2,

Figure 2

Maturation of semiotic systems over time.



where the growing area of the transverse plane represents the increasing scope of human capacity from infancy progressing toward maturity or, in general, from not knowing to knowing.

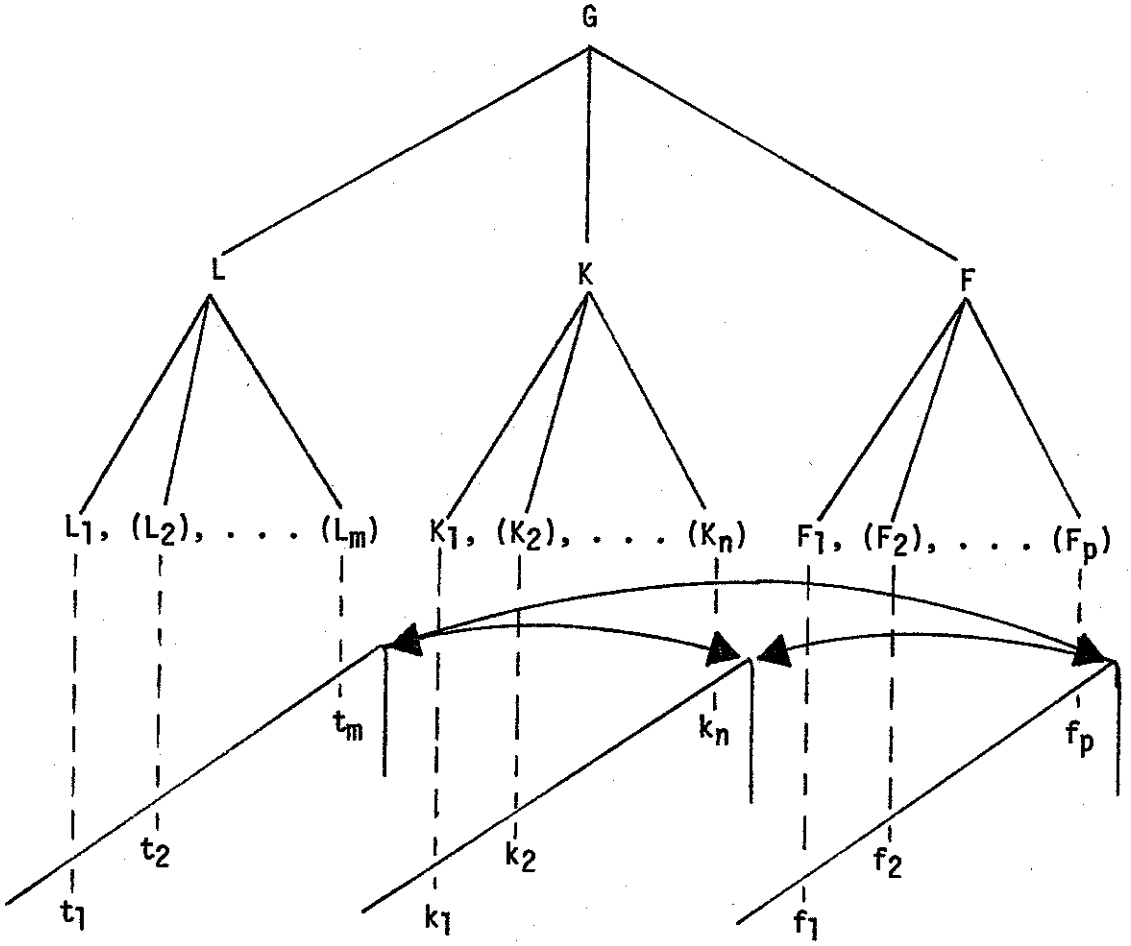
But, as will be obvious, the interrelationship of texts and facts, as theorized in Figure 1, developing over time as in Figure 2 requires elaboration. In actuality we do not find a simple dichotomy. Rather, we find something more like the differentiated hierarchy of semiotic systems suggested in Figure 3. As intelligence (alias communicative competence) develops, the systems of the hierarchy become both more thoroughly differentiated at the surface and more completely integrated at deeper levels. In Figure 3, three types of semiotic systems are posited, with the middle term hypothesized to be actually intermediate to the other two. Building on and expanding on the simple theory of Figure 1, at the extreme right are posited semiotic systems for the interpretation of "facts," designated F_1, F_2, . . . F_p (where p is some number), and dominated by a more or less universal and largely innate logic (universal grammar) of facts, designated simply F. At the extreme left of the figure, a similar configuration of systems is shown for the interpretation of texts expressed in particular languages, L_1, L_2, . . . L_m, dominated by a system of universal grammar designated L, again, hypothesized to be largely innate. Intermediate between these two types of semiotic systems is yet another series, K_1, K_2, . . . K_n, kinesic or gestural systems that are more or less coordinated with particular systems of facts, F_i, and/or languages, L_i. Again it is supposed that the particular K_i are dominated by a universal system of gestures, K, also hypothesized to be largely innate.

Subtending the hierarchy of L_i, K_i, and F_i will be particular elements of the sensorimotor system (actually a subhierarchy of systems) used for interacting with the world of experience through sensorimotor activity and also for translating among the three types of semiotic systems and thus integrating them at the sensorimotor level. For instance, hearing or uttering a text in a particular language will depend on certain auditory or articulatory systems that are integrated with the grammar of a particular L_i, while reading or writing the same text will depend on visual or manual systems similarly linked up to the grammar of L_i. Similarly, K_i will be subtended by particular manifestations in the form of kinesic "texts" and at the same time will be integrated with both F_i and L_i. By the same token, F_i will be subtended by particular sensorimotor systems and at the same time integrated with the other types of semiotic systems. The three types of semiotic systems (and whatever others it may be necessary to postulate in a more complete theory) are themselves dominated by a general semiotic system designated G, for general intelligence or general communicative competence.

The model is founded on the premise that all modes of interacting with the facts of experience are dependent directly on more or less veridical representations of the facts themselves, as given by innately structured sensorimotor subsystems, or on other representations that, to some extent, at least, may be translated into sensorimotor representations. Further, it is postulated that imaginations (fantasies and the like) are factual and may be factually represented, but are not themselves veridical representations of external facts, and it is postulated that the basic problem of intelligence is to differentiate factual representations (true, or well-calibrated texts) from nonfactual (false, or poorly equilibrated texts). C. S. Peirce (1877) described this problem as the resolution of doubt; Dewey (1916), as the resolution of trouble or the discernment and resolution of meaningful conflicts, for example, trying to

Figure 3

A slightly expanded conception of pragmatic mapping with a gestural semiotic system (K) as intermediate to universal grammar (L), and to a universal logic of facts (F).


G
L
K
F
L_1, (L_2), . . . (L_m)
K_1, (K_2), . . . (K_n)
F_1, (F_2), . . . (F_p)
t_m
k_n
f_p
t_2
k_2
f_2
t_1
k_1
f_1

stay alive. Probably some such overall guiding plan--additionally incorporating the avoidance of pain and the seeking of comfort, avoidance of anything that damages or injures the self-concept, or altruistically avoiding things that damage the concept(s) of valued other(s) and seeking that which enhances the self and enriches valued other(s)--will be necessary to any adequate theory of ordinary intelligence and equally to a theory of general communicative competence.

Figure 4 elaborates only slightly on the sort of integration envisioned between the various representational/interactive (semiotic) systems that are postulated in the pragmatic theory under consideration. The conceptualization envisioned is less a theory about "hardware" (brain circuitry) than it is about "software" or programming of the general hierarchy of semiotic systems.

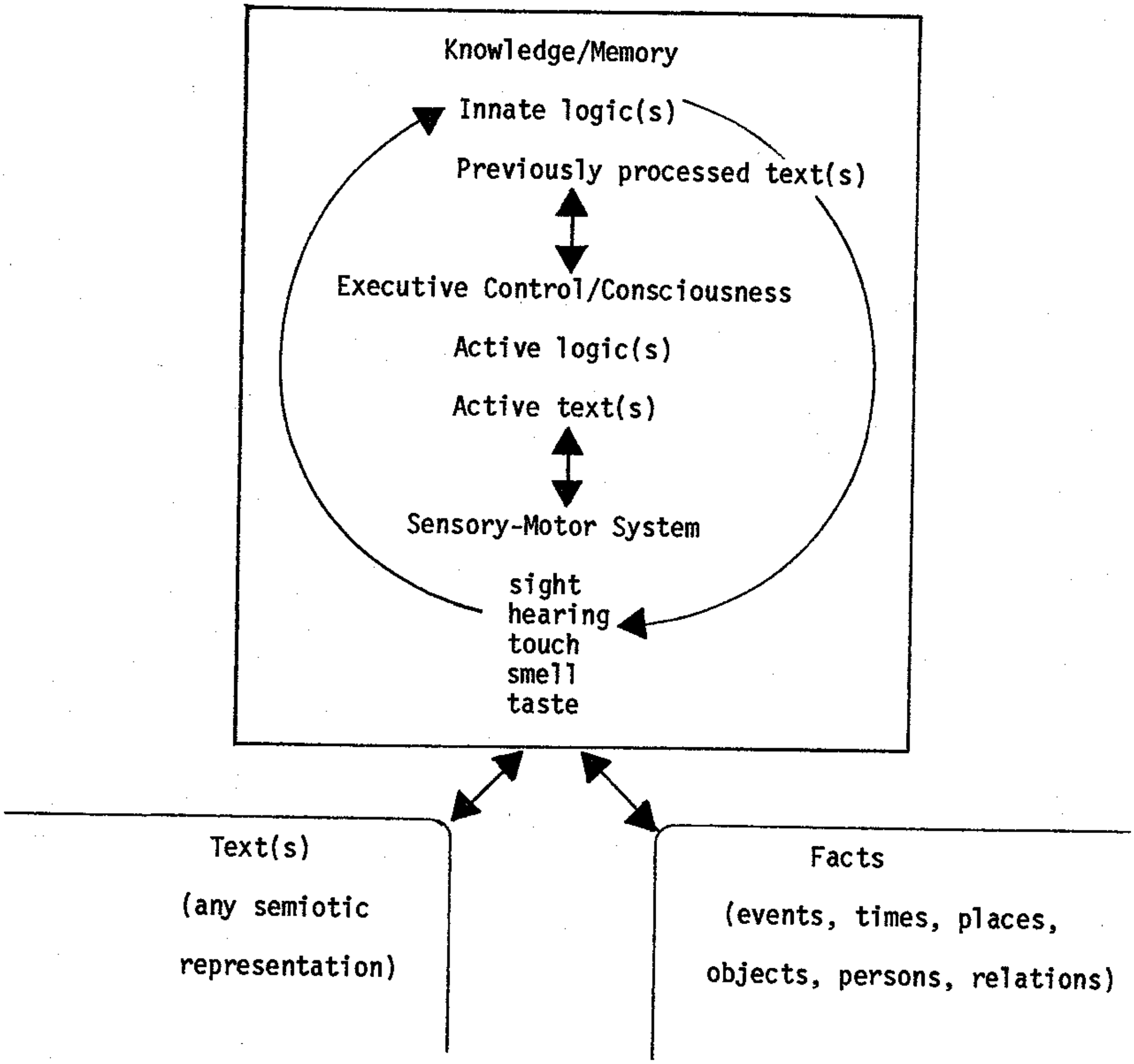
There are really two points of interaction with externality: On the one hand, there is contact with the real world of facts through the sensorimotor systems and, on the other hand, there is contact through the creation and manipulation of abstract texts in the various semiotic systems the person possesses innately and/or develops or acquires over time. With respect to actual public texts, those that are realized as elements in the world of facts, the operations are largely overt and may be intended for interpersonal communication, but in the case of private texts that may or may not be realized in the world of facts, the operations may be largely covert and may be used chiefly for intrapersonal communication.

Intermediate between the texts of L_i and the texts of F_i are those of K_i. Before turning to some explicit empirical hypotheses about how these systems interact over the course of maturation (including primary or nonprimary language acquisition) and how the various elements relate within the overall hierarchy, it may be useful to examine the proposed integration of the hypothesized modular components within the system and to see how all of them together offer a reasonable basis for characterizing the interrelatedness of foundational mental constructs such as consciousness, memory, knowledge, belief, volition (intention), and so forth.

Since one of the aims of this conference is to evaluate the extent to which TOEFL adequately assesses the construct of communicative competence, it may be useful to begin by asking how competence itself is defined within the conceived theory. Actually, competence in the desired sense is precisely analogous to what is commonly referred to as knowledge. It comes into the model at every level of programming and interpretation. There are basically two types of knowledge--innate or prewired and acquired or interactively devised. No one really doubts that it is necessary to hypothesize, as Chomsky (1965, 1975, 1980, 1982) has amply demonstrated (also see Lightfoot, 1982), that the human being is preprogrammed with certain kinds of innate knowledge, though there is a great deal of disagreement about just what aspects of knowledge must be innate.

Figure 4

A slightly elaborated model of the interrelation of various mental constructs as components of general intelligence, alias communicative competence.



In the model proposed here, innate knowledge is postulated as the basis for three distinct types of semiotic systems, L, K, and F, and underlying all of these, a general semiotic, G, referred to simply as *general intelligence* (or, alternatively, general communicative competence). In addition to this, the sensorimotor systems also represent innate knowledge. For example, as Piaget (1947) so aptly argued, the sucking response in infants presupposes the shape of the breast, the nutritious value of the milk, the entire digestive process, and a great deal more. The very shape of the mouth and breast are co-adapted for certain ends. The same can be said of physical systems in general in biological organisms.

In addition to innate knowledge, there is knowledge that is somehow the product of the interaction of innate knowledge with *experience*. Within the proposed theory, the latter (experience) is defined as *the temporally progressive textualization of facts and events that impinge on the organism*. The nature of this textualization (interpretive or creative interaction) is determined by the character of the innate hierarchical system of semiotics. The organism, in other words, is free precisely within the limits of the semiotic systems it innately possesses. Beyond those limits it cannot go, but it develops within them. Without innate knowledge in the sense argued by Chomsky, agreeing with Rene Descartes, Immanuel Kant, and Albert Einstein, experience itself would be impossible--literally unknowable, unutterable, and unthinkable. However, given the a priori knowledge constituted by the innate semiotic hierarchy as postulated by Chomsky, experience becomes possible and, with it, additional knowledge. That is, as texts are interpreted and created through the interaction that we call experience, new knowledge is acquired. While this knowledge is being acquired, we refer to the whole experiential process as *consciousness*, afterward as *memory*. Thus, together, *consciousness* and *memory* are closely analogous to the linguist's term *performance*, while *knowledge* is analogous to *competence*.

It is necessary, however, to distinguish degrees of consciousness. One is not exactly unconscious during sleep, but within the model (see Figure 4) it is quite convenient to distinguish between *wakefulness* and *sleep* by referring to the degree to which the sensorimotor system (and the rest of the rational apparatus) is on or off. Waking up, within the proposed theory, may be defined as arousal, especially of the sensorimotor system, and *alertness* as the extent of that arousal. Beyond alertness we may define *anxiety* as an overload arousal that, if prolonged, may lead to various pathological conditions, whether short-term, such as "being paralyzed with fear," or long-term, such as what is popularly referred to as "nervous collapse."

The proposed theoretical framework suggests that communicative competence, or intelligence, is *a dynamic state that obtains within a hierarchy of semiotic systems to the extent that those systems are well equilibrated with reference to each other and to externality*. The dynamic, active character of the hierarchy is necessary in order for it to maintain or to seek equilibrium. Further, we may refine the definition of *communication* by saying that it is *activity within the system aimed at achieving or maintaining equilibrium*. In fact, as Peirce (1877, 1878, 1905), Dewey (1916), and others observed long ago, conscious intelligent activity is almost always aimed at resolving disequilibrium, doubt, conflict, trouble, and the like. The intellect gets going and we start talking, thinking, communicating, just at the point where the conflict (in the sense of disequilibrium) arises.

In addition to the mental constructs already discussed, it is necessary to add *beliefs* and *need states* (including biological drives, innate psychosocial needs and tendencies, and all other sorts of semiotically defined goals). A *belief* may be defined, within the proposed theory, as *a textual representation (a simple proposition or complex of them) that is taken to be true* (in the loose sense of being *more or less equilibrated with some corresponding set of facts*). Within the theory, *orders of belief* are distinguished in terms of the degree of certainty associated with them, and *classes of belief* are distinguished on the basis of the distinct semiotic bases that are possible, that is, *classes of semiotic representation*.

First, we will consider *orders of belief*, then *classes*.

At least three orders may profitably be distinguished within the model. *First order beliefs* are those arrived at through experience, the sort we take note of when we say "seeing is believing," or "I saw it with my own eyes," "I tell you I was there," and the like. Experience is taken as prima facie evidence of the truth of whatever facts are (or seem to be) represented by it. However, this sort of belief is *not*, according to the proposed theory, the most basic.

Zero order beliefs are those textual representations accepted as true (in the defined sense) because they follow from some logic regarded as necessary or because they are given by one or another aspect of the innate system (e.g., chocolate cake tastes good because it tastes good; water is satisfying when one is thirsty; sex is interesting because it is interesting). Intermediate between zero order and first order beliefs are all those representations normally taken for granted in a given individual's experience but that may, under certain circumstances, be drawn into question--e.g., is the desk really hard? Is death a necessary eventuality? Do organisms cease to exist in all senses when they die physically? Are there mental powers that cannot normally be perceived and that are not limited by spatiotemporal physics?

Progressing away from absolute certainty (or whatever may be its closest actual approximation in any given human psyche) beyond experience, we come to a vast array of representations clamoring for a hearing. Among these, within the proposed theory, texts accepted as true on the grounds of the first order beliefs of others (that is, on the basis of reported experience or testimony) are *second order beliefs*. Beyond these it is theoretically possible, though perhaps less useful, to distinguish additional orders of belief--*third order belief* corresponding to acceptance of some representation(s) by a third person concerning the experience of a second person, and so on progressing to more remote orders.

In addition to *orders of belief*, it is useful to distinguish *classes of representation* within the proposed model and corresponding *classes of belief*. While the order of a belief is an index of the a priori degree of confidence associated with it, its representational class partially determines how the confidence per se is achieved. As noted above, some representations are regarded as true because they seem to arise naturally in experience. The sun looks bright; therefore, the sun is bright. Thus some zero order beliefs are engendered by sensorimotor experience by virtue of the natural (innate) correspondence between the sensorimotor system itself and the facts (or for skeptics, the apparent facts) of externality. Thus, the most superficial level of semiotic representation, and the class of representation that appears to be closest to actual facts, is *the sensorimotor class*.

Sensorimotor representations not only give rise to certain zero order beliefs, but they also enender first order beliefs. They are, themselves, in some unquestionable sense assumed to be actual or real. This assumption may be naive, but it is not trivial, and generally it is correct. Although sensorimotor representations may or may not be well equilibrated to facts, they are generally believed to be true (sometimes falsely, as in the case of illusions and hallucinations) precisely *because* they are semiotically represented by the sensorimotor system and are generally unaffected by volition. That is, they are not ordinarily under conscious control (hypnosis or mere "wishful thinking" being special cases where such representations may indeed come under one's own or someone else's conscious control).

In addition to sensorimotor representations, there are intentional, *volitional*, and nonintentional, *nonvolitional representations* that may or may not be translated (or translatable) into the sensorimotor class. Volitional/nonvolitional representations are not completely mutually exclusive, but rather form a continuum based on greater or lesser degrees of conscious control. Representations that come more definitely under the control of consciousness, *volitional* (or intentional) representations, are generally termed thinking, reasoning, planning, problem solving, listening, speaking, reading, writing, and the like. However, fantasizing and imagining may also to some degree come under conscious control. The degree of truth associated with such representations, that is the extent to which they are believable, is a function of the extent to which they can be set in correspondence to representations of the sensorimotor class, that is, the extent to which the distinct classes are intertranslatable. The nonintentional class of representations would include dreams at the extreme opposite end of the intentional and nonintentional continuum, which merge with imagining, daydreaming, and fantasizing, extending toward the opposite end of the scale that would be occupied by clearly intentional representations and volitional acts.

An important element of the theory proposed here is the postulation that all organismic activities, including purely mental events and therefore all representations, are in themselves necessarily real. That is, they are facts. Therefore, there may exist true (i.e., correct and appropriate) representations of them. However, the major problem of intelligence is to determine which representations qua representations correspond correctly and appropriately to these and other facts. Or, more explicitly in the terms of the model, the basic question is _which_ abstract representations can be regarded as true in the sense of _being_ well equilibrated (or intertranslatable) with sensorimotor representations. The fundamental problem of intelligence, therefore, is to _equilibrate_ representations so that they fit the facts--to determine what is to be believed and, therefore, acted upon. Or, putting the same problem differently, it is to know what the facts are--how they may be interpreted.

The model takes for granted that beliefs may be upgraded or downgraded by obtaining corroborating evidence or testimony, by confirming or disconfirming the reliability of the witness(es), by scientific experimentation, by logical reasoning, and so on. The idea of absolute equilibration--absolute truth--is inescapable here, but the conclusion that such a possibility logically is implied by the theory is not evidence for or against its existence. It is, in some sense, an a priori argument in favor of the logical necessity of absolute truth, but the proposed theory is neutral on the question of whether or not human intelligence can know such truth. Though it is a simpler matter to develop a case against the human attainability of any absolute equilibration than to develop one in favor of such a thing, it is important to realize that such arguments cannot touch the possibility of whether or not such truth exists (indeed they must presuppose this entirely logical possibility in order for the arguments themselves to occur in the first place).

Passing on to the matter of _need states_ and their place in the model, we may begin by noting that biologically defined need states (drives, mental requirements, or mere proclivities and tendencies) are roughly on a par with unquestioned representations (zero order beliefs). They are representations that constitute _zero order plans_ for interpreting or adjusting (acting upon) the facts of experience. As such, they correspond roughly to Freud's concept of the id. These zero order plans are also reflected, as Freud observed, in dreams. Representations that come under conscious control, however (including _first_ and _second order plans_), would correspond roughly to Freud's concept of the ego. Those plans that come to be regarded as "true" or right, by the highest principles of reasoning and by the highest standards of equilibrium (whatever is regarded as ethical, or moral), would correspond to his idea of the superego.

To the extent that the results of higher reasoning (above the representations provided by sensorimotor experience) lead to what the person regards as true representations, and to the extent that the person comes to act on these, automatically as a matter of course it may be argued that they have become incorporated as zero order plans. Probably only zero order plans (those incorporated into the subconscious), first order plans (those consciously arrived at in ongoing experience), and second order plans (those derived from the reported experience of other persons) will need to be taken into account in a theory of communicative behavior, though a theory of competence will probably need to go further.

IV. Advantages of the Theory

In this section, certain questions about the theory are answered, some hypotheses are suggested, and an application to a problem in kinesics is discussed.

A. Arguments on behalf of the model

1. Why does the model implicitly accept Chomsky's innateness hypothesis?

We know that the physical world interacting with the physiological mechanism defines certain a priori limits of possible semiotic systems. There is a rapidly growing body of evidence showing that unless the human mental apparatus is preprogrammed with certain highly abstract and very complex notions about how things in externality are to be interpreted, it would be quite impossible for us to acquire natural languages. The arguments along this line (see Lightfoot, 1982) are profoundly persuasive to those who examine them seriously. The evidence on this question is not yet sufficient to flesh out the details of a biological basis for universal grammar, but it is undeniably adequate to show that some such basis must exist and to suggest general outlines of the system.

2. Why does the model start with pragmatic mapping?

The justification for this decision is purely heuristic. However, it is motivated by the likes of epistemologists such as C. S. Peirce, John Dewey, William James, Jean Piaget, Ferdinand de Saussure, and Albert Einstein. The renowned physicist put it this way: "Everything," he said, "depends on the degree to which words and word-combinations correspond to the world of impression" (1941, p. 112). If the term "words" is generalized to include all sorts of semiotic representations ("texts" in the most general sense), and "the world of impression" is interpreted to mean the world of experience, we do no violence to Einstein's intention and we arrive at the essential basis of the entire concept of pragmatic mapping as a fitting or equilibrating of texts to facts and this as the central activity of human intelligence.

3. But why a unified model?

The basis for proposing a unified model is an empirical one. In order for human beings to talk about what they see or to visualize what is talked about, or heard, or touched, or imagined, or, in general, to translate between distinct modes of semiotic representation, there must be a basis for the integration of all the distinct representational systems (see Fodor, 1975, 1980). Moreover, if there were no such system of integration, it is inconceivable that human beings should have even so much as an illusion of an integrated stream of experience. Further, communicative competence surely depends in some measure on the ability to achieve the sort of integration between distinct semiotic systems as suggested by the model. Therefore, anything less than a unified model will miss each and every one of the marks defined by goals 1-6 above. That is to say, models that examine isolated components of the proposed hierarchy will fail to achieve an adequate conception of communicative competence or of how to test it or to evaluate tests of it. Also, the proposed theory agrees in spirit with the requirement of Dan Douglas (this volume) and others that tests must be "domain specific." More on this below.

B. Some general hypotheses suggested by the model

1. The dominance hypothesis

The development of any given semiotic component (system or part of a system) beyond its initial state will depend on the development and differentiation of subordinant semiotic systems. That is to say, general intelligence cannot be expected to advance normally except to the extent that the semiotic subsystems L, K, and F develop and are differentiated. In other words, to the extent that the development of the semiotic systems mediating between sensorimotor representations and general semiotic representations is retarded, intelligence will be retarded, and conversely, the extent to which intelligence is organically retarded will set limits on the possible growth and development of the intermediate systems and thus on the understanding of sensorimotor representations.

2. The primary language hypothesis

The single most important factor governing the development of general intelligence in normal human beings will be the development of skill in a primary language (not necessarily including speech, however). To the extent that this development is retarded, general intelligence will be correspondingly limited. However, it is possible that development begun in one language (say, Italian) will be continued in another (say, French) and that development in yet some other language (say, American Sign Language, or English, or something else) may end up being the "primary" or main language. This sort of development will not necessarily impede the normal growth of intelligence, but, just to the extent that the subject is prevented from attaining optimal growth in a (any) primary language, the normal growth of intelligence will be impeded. Conversely, to the extent that growth in a primary language is facilitated, the optimal development of intelligence will be ensured. Perhaps this hypothesis will prove incomplete and will have to be expanded to include the development of other semiotic systems on a par with natural language developments, but I know of no evidence to support this possibility. On the other hand, there are cases such as Genie's (see Curtiss, 1977) that support the hypothesis as stated.

3. The contiguity hypothesis

Transfer across systems will be greatest for those that are dominated by a common node in the model and may be expected to spill over to indefinitely distant systems via mediating elements of the hierarchy. That is to say, for example, acquiring a second language will be influenced greatly by the prior acquisition of the first. Or, to take another example, acquisition of skill in reading may be expected to have a positive impact on speaking ability in the same language. Taking up tennis (an example of an F_i mediated by particular sensorimotor skills) can be expected to have more impact on a closely related skill such as racquetball, but it will also potentially impact (and be affected by) something more distant, such as karate. However, it is supposed that the contiguity hypothesis will result in observable effects only if the skill-ratio/threshold requirement is met.

4. The skill-ratio/threshold hypothesis

Transfer is not expected to be observable unless the relative maturity of two contiguous systems is such that one system is markedly more mature than the other. For instance, the acquisition of literacy is expected to impact oracy noticeably only at the point where reading skills significantly surpass speaking skills. Stated generally, the acquisition of an _i_ + 1 system will have little or no observable impact on the _i_th system until skill in _i_ + 1 significantly surpasses skill in the _i_th system.

5. The correlation of skills hypothesis

Contiguous skills, in general, are expected to correlate with each other more strongly than noncontiguous ones, but this correlation is expected to be modulated by the relative maturity of the respective systems where equally mature systems will be optimally correlated. For example, _ceteris paribus_, to the extent that one has had a well-rounded educational experience, oracy should be optimally correlated with literacy in one's primary language. However, there is less reason to expect to find such a correlation when one's experience has been such as to produce much more emphasis on the development of oracy than literacy. Or, to take a different example, one's ability to perform closely related motor tasks may be expected to correlate more strongly than one's ability to perform unrelated tasks, and, again, the correlation will be modulated by the relative opportunities to practice and develop skills in the two sorts of tasks.

C. An application of the model in the gestural domain

How does a particular gesture come to be indexed as having a particular range of meaning within a given system of kinesics (K_i)? For instance, how is it that a brandished fist may come to mean either a threat to punch someone in the nose or a symbol of solidarity or brotherhood? Or, to take a different example, how is it that in one culture a gesture with the index finger palm upward means "come here," while in another culture, the same meaning is coded with roughly the same gesture palm downward? Also, just what aspects of the meaning assignments for such gestures are innate and what aspects are acquired? Answers to these questions will not, of course, prove the model, but they will help to illustrate how it might in fact be applied to some common empirical problems, and it may help readers to see at least how the model may be interpreted so as to be made vulnerable to close empirical scrutiny.

Take the brandished fist as a threat. The model suggests that the development of an interpretation of a brandished fist as a semiotic (kinesic) representation of a threat to punch someone in the nose has its roots in a general, and probably universal, kinesic system linked up with sensorimotor experience. At the deepest level, the meaningfulness of a collision of two objects is a propositional relationship between two arguments (probably an actor and a patient) in relation to a given predicate (the momentum of one of the arguments as contrasted with the inertness of the other). The observable result that the collision may cause damage to one or both arguments of the momentum predicate is an empirically discoverable fact, given the innate logic that enables the experiencer to have such an experience in the first place. Without such an innate (general semiotic) knowledge about facts, it is impossible for the experiencer to have the experience of two objects colliding or to have any understanding of the nature of a collision.

At the same time, the innate knowledge is apparently only accessible as it is called into play in the interpretation of actual experience. When this happens, the innate knowledge makes possible the initial interpretation of the facts and its own subsequent enrichment as a result of memory. That is, the fact that colliding objects may damage each other is noted and may be subsequently recalled without having to reexperience a particular episode of collision. Of course, the brandished fist implies something more, especially the intentional agent-hood on the part of the person who waves the clenched fist, and potential patient-hood on the part of the person at whom the fist is waved. The innate system of logic must supply such categories as "agent" and "patient" to differentiate potential arguments of particular predicates. Given such categories in an innate logic, the possibility of applying them to interpret a brandished fist as a threat--namely a prelude to a punch in the nose--is a straightforward semiotic extension. It also shows how the semiotic systems of gestures (K_i) in the model are intermediate to systems that pertain to facts in general (F_i) and those that pertain to languages (L_i).

The extension of the same gesture to a seemingly contradictory meaning--namely, "solidarity"--can be achieved in a variety of ways. For instance, in spite of the arbitrary conventionality of the connection between the symbol and this, perhaps secondary, meaning, a certain family resemblance to the primary meaning may nonetheless be discerned. More than one person can be construed as the agent, attacking one or more other persons--uniting against a common enemy to deliver a punch in the nose. Or, another possibility is to understand the meaning of the gesture as drawing persons together the way one's fingers draw together in making a fist. In either case, the conventional interpretation may be "fixed" (that is, determined, or made secure) by associating it with a factual context (i.e., a "text" of some F_i) such that the interpretation of the brandished fist as meaning "solidarity" is well equilibrated. That is, it fits the facts, while the meaning, "I'm going to punch you in the nose," does not fit.

Thus, the innate logic of general semiotics can be shown to apply to particular gestural interpretations (in a given K_i) and at the same time to the general logic of factual occurrences $(F)^i$. Also, it is not difficult to extend the same reasoning to show how particular elements of a gestural system would by the same sort of translation be related to one or more languages (L_i) via the intermediating grammars of those particular languages. This sort of connection would conveniently explain the fact that a palm down index-finger-wagging gesture might mean "come here" as used by the speakers of one language in a particular context, while a similar gesture palm up might be the equivalent in another culture. Or, alternatively, it would explain why both gestures might have completely distinct meanings in some third culture.

VI. Implications for Tests

It remains to be shown that there are explicit practical implications of the proposed theory for tests of comunicative competence and that, to some extent, these differ from the implications of competing models. I will argue that the proposed model results in a more practical set of criteria for tests of communicative competence than do those derived from the notional/functional perspective that uses the dyadic definition of communication as its implicit starting point. Further, it yields an explanation of empirical findings in testing research that remain largely unexplained by taxonomic approaches. First, we will examine criteria for tests of communicative competence and, then, empirical findings in testing research.

A. Criteria for tests of communicative competence

1. From the taxonomic perspective

Here I propose only to examine the perspective put forward by Andrew Harrison (1983), who chooses to work within what he understands as the notional/functional paradigm. His approach is useful because of its explicitness and because he attempts to lay down clearly the limits of communicative testing. However, I do not claim that his views are common to other theorists who have been influenced by the notional/functional paradigm. In fact, it is clear from the very excellent paper by Duran et al. (1985) that, although they are offering an admittedly taxonomic approach, they would prefer a more coherent, generative theory.[2]

The objective here, however, is limited to showing that Harrison's criteria, based on the dyadic definition of communication, are too confining. Beyond that, any taxonomy that fails to account for the integration of its hypothesized components will eventually need to be modified into some more coherent system. In the meantime, I think that top-down reasoning, from theory to data, will in some ways be as profitable as bottom-up reasoning, from data to theory. For reasons that are becoming clearer, a modular theory (see Chomsky, 1982, and Lightfoot, 1982) will probably be required. My expansion of this idea suggests that the modules will need to be arranged in an integrated hierarchy with general intelligence (communicative competence) at the top of the tree and various sorts of representations (texts) at the extreme surface (ends of branches).

In any case, Harrison's criteria are as follows:

a. "A test should assess language used for a purpose beyond itself" (p. 77).

b. There must be "an information gap" such that the pressure on both interlocutors to "know or tell" is similar. (p. 77)

c. The encounter must result in changes in the interlocutors. (p. 78)

On the basis of these criteria, Harrison rejects "some types of tests which are currently advocated and widely used, such as cloze and dictation" because, he says, "they do not involve the student in any useful activity" (p. 78). Harrison contends, "A test type does not become communicative by mixing in a dash of reality: it is communicative because of the use made of it, and if it cannot be used to represent communicative purpose, it cannot be a communicative test." (p. 79). So, Harrison argues, if the test does not involve at least a pair of interlocutors changing each other by virtue of some shared purpose beyond the test per se, that is, reducing an information gap, the test is not a communicative one. In all of this, it is clear that Harrison (who, according to his own assessment, is an interpreter of the notional/functional paradigm, [p. 79]), assumes the standard dyadic definition of communication.

2. Criteria from the more general definition

From the more general definition, different criteria follow:

a. A test of communicative competence is an intelligible representation (text) in some semiotic system.

b. An individual's communicative competence with respect to that text may be construed as the degree of intelligibility of that text to that individual.

c. The validity of a particular text as a test of communicative competence will be limited by the extent to which it engages and effectively challenges the intelligence of the examinee attempting to produce or understand it.

The first criterion derived from the general definition of communication and the corollary definition of communicative competence includes *any* test that engages intelligence to any degree as a measure of communicative competence. The second criterion, however, takes account of the fact that a given test performance is just that--performance of an examinee on that specific individual test. If this criterion is understood, together with the third, it takes account of the important concerns voiced by Douglas (this volume) regarding the need for domain specificity. However, we need to realize that *the factual domain also includes the texts that are typical of that domain and the performances of typical persons in the utilization of such texts*. The third criterion adds the all-important element of validity. It suggests that the overall validity of a test is to be judged by the extent to which examinees on the whole *try* to render intelligible whatever text(s) the test consists of or results in the production of. At the individual level, the same interpretation applies. The validity of the individual's score will depend on the extent to which the test engages and challenges that person's ability to perform that particular task. (The performance of an individual who remains indifferent or preoccupied with something else throughout the taking of a test cannot necessarily be taken seriously as an index of any sort of communicative competence.)

It might appear that the criteria of the general definition suggest that all tests are equally good or bad measures of all sorts of communicative competence. This, however, is not the case. All tests are *not* equally effective at engaging and challenging the individual's capacities (the third criterion), and, moreover, the second criterion makes explicit the fact that different texts offer different indices of communicative competence. Furthermore, if the criteria are taken together with the entire scope of the pragmatic theory presented above, it is clear that distinct elements of the unified hierarchy of semiotic systems may be challenged differentially by different tests. That is to say, the theory allows for trait variance *and* method variance. Although it makes testable claims concerning ways in which the hierarchy is differentiated, it also makes testable claims concerning the particular manner in which elements of the hierarchy are integrated. The model as proposed has theoretical advantages in terms of its generative character and its modularity.

Also, it allows for a straightforward elaboration of the concept of validity in terms of the process of pragmatic mapping. This theoretical concept is applicable in more than one way. The theory adopted here applies to the linking of texts (semiotic representations) to meanings (other semiotic representations), some of which are ultimately tied to facts in the world of experience. Since it aims to distinguish those tied to facts from those that are not, the theory of pragmatic mapping includes and extends beyond the basic research question Chomsky posed for the investigation of natural languages early in the development of generative theory--the problem of the correspondence between sound and meaning.

At a higher remove--a more abstract and more general level of investigation--pragmatic mapping applies to the problem of differentiating degrees to which semiotic representations are well equilibrated to the facts of experience. That is, the theory of pragmatic mapping defines the general problem of scientific investigation--as Einstein put it, the correspondence between "words and word-combinations" and "the world of impressions." The equilibrating of that correspondence, chiefly by adjusting the semiotic representations until they optimally fit the facts--explain, predict, control, and so on--is the scientific enterprise.

And, in a more specific way, the same concept of pragmatic mapping and the equilibrating of semiotic representations defines the problem of test validation. Our task is to determine when a given performance faithfully represents a person's ability. Part of the problem is to define the factual domain of the performance (the facts side of Figure 1 above) and part is to determine what sorts of texts to use (the other side of Figure 1). But more importantly, the test validity problem requires determining the degree to which the particular semiotic tasks selected for the test actually reflect the normal, typical, and actual relationships between texts and facts in the determined domain.

Frankly, I do not think that this problem can be solved if the domain of facts remains unanchored (unspecified, determined). It must be fixed. This is at least part of the job of a theory of communicative competence or of pragmatics. Actually, at least some of the notional/functional approaches to the problem have bogged down precisely because they have failed to fix the factual domain. An example contrasting cases where the facts are fixed and where they are undetermined will help show what I am getting at.

Solidly within the tradition of the notional/functional syllabus, Jones (1979) suggests that students in an ESL class should make a list of five things to ask the teacher's permission to do. He tells them:

> Sometimes we need to do more than just offer to do something--we may need to ask permission to make sure we are allowed to do it. The expression to use depends on: (a) the type of task you want to do and the degree of resistance you anticipate; (b) who you are and who you are talking to--the role you are playing and your status. (p. 223)

I have put this sort of task to native speakers of English, and they are invariably nonplussed. Why? Because the facts are insufficiently determined. Ordinarily we don't go up to anyone to ask permission for even one thing out of the blue, much less for five. There needs to be an experiential basis for doing so. That is, there needs to be a particular reason for us to ask permission, and the other person must have some basis for regulating the proposed action.

If the problem is difficult for native speakers of English, what is it for nonnatives? Imagine being instructed in Chinese, or some other language that you do not know well, to perform the sort of task that Jones proposes. The trouble for the nonnative is that not only are the facts undetermined, the needed textual forms too are largely unknown--and for most nonprimary language acquirers, the instructions for the problem will be incomprehensible. The problem is something like linking up two spaceships, both of unknown dimensions, somewhere in outer space at an unknown time.

If we happen to be native speakers of the language in question, one side of the pragmatic equation (Figure 1), the text side, is more or less determined. The problem is solvable by using known textual forms to determine the facts. If the facts were given, the problem would be relatively easy for the native. However, for the nonnative, if neither the text(s) nor the facts are given, the problem is essentially unsolvable. (Indeed, one side or the other must be fixed, or the problem is essentially unsolvable for anyone!) However, if the facts are fixed (determined and interpreted), even the nonnative has the potential of comprehending and eventually producing the appropriate sorts of texts.

The generalizations to testing are straightforward, but perhaps not obvious. We will look at them specifically in a moment, but first, I think it may be helpful to push the point one step further. We need to look at an example where the facts <u>are</u> fixed and see how it contrasts with the other case. Consider an incident in <u>La Familia Fernandez</u> (a program for teaching Spanish as a foreign language: Oller, 1963, 1965).

Pepito, a five-year-old boy, is kneeling in the yard and petting his dog, Iman, when Enrique, a teenager, arrives looking for Emilio, who is Pepito's teenage brother. Enrique greets Pepito and asks for Emilio. At just that moment Pepito looks back over his shoulder and points out that Emilio is coming out of the house. Emilio and Enrique exchange greetings. Enrique asks if Emilio is ready. The latter says he is, holds up his bag in a gesture to demonstrate that he is ready to go, and the two of them start out for the gate. Pepito follows. So ends the first episode of the story. Then, in the next, Pepito is shown catching up with the older boys and asking where they are going. Emilio tells him it's none of his business. Pepito appeals to higher authority. He yells for Mom. Emilio tells him to quiet down. Pepito asks again where the two older boys are going. Emilio relents and tells him that he and Enrique are going swimming. Enrique agrees in a conciliatory manner. Pepito asks permission to go along. Denied. More yelling for Mom. Emilio tells him to be quiet again and looks desperately at Enrique, who nods reluctantly; Emilio says Pepito can go along. Pepito changes his mind, now that he has won.

He says, "No. I don't want to go. I'm gonna play with Iman." He returns to the yard, running after the little terrier.

What does this lesson have that Jones' lesson lacks? A world of experience. The facts are fixed in the story about Pepito, they are not in the other case. For instance, there are clear and determinate answers to questions regarding the latter case while there are none in the former. Why does the student want the teacher's permission for the five things? For any single thing? Why does the teacher want to regulate the proposed activities? Who cares? I am not saying that these questions are absolutely devoid of answers with respect to any specific request once it is stated (that is, with respect to any particular act of pretending to ask permission); what I am saying is that the answers to the questions are indeterminate. They must be invented. There are no determined facts that would make the answers to the questions accessible to reason. Reason must invent them.

But, compare. Why does Pepito want to know where his brother and Enrique are going? Why does Emilio tell Pepito to mind his own business? Why does Pepito yell for mother? Why does Emilio tell him to be quiet? Why does Emilio finally give in? How come Pepito changes his mind? Is this plausible? There are, in fact, literally, in fact, answers to the questions about the Pepito story because the facts have reality within the context (albeit a fictional one) of the story.

Now for the generalizations to testing, in brief. To the extent that the facts remain unfixed in any given communicative activity, just to that extent will it be difficult to meaningfully evaluate any individual's performance in that activity. The activity, or task, will be difficult to scale in a meaningful way. Further, the validity of any test based on it will be indeterminate. On the other hand, to the extent that the facts are fixed in any communicative activity, and to the extent that the criteria mentioned above are met, the task will be scalable in a meaningful way. Or, putting the whole matter in a slightly different way (but still within the same pragmatic perspective), meaningful context is a dynamic equilibrium betwen some representation(s) and facts. Following the same logic, an activity's authenticity is interpreted as the degree of pragmatic equilibration between text and facts. The dimension of simplicity or complexity will be meaningful only to the extent that the factual domain of text(s) accessed by the testing process is fixed by it, that is, the degree to which the text(s) may be judged authentic or well equilibrated to an established, determinate factual domain.

To call for domain-specific testing, as Dan Douglas does (this conference), is to insist that the factual domain be more fully specified. This is the sort of "anchoring" of the factual domain that most of us are probably in favor of. To the extent that it is accomplished, construct validity of our tests will be guaranteed, and at the same time, criterion-validity (in the sense of Popham, 1975, 1978, 1981) will equally be assured. In just this way, the meaningful scalability of any test depends on fixing the factual domain. A pragmatic theory is required for this purpose and, when accepted, thus converts all testing into specific-purpose testing (and, to Popham's sense of criterion-referenced testing). Indeed, to the extent that testing is pragmatic in the required sense, it is "specific-purpose" oriented and criterion-referenced.

B. Explaining the empirical findings

It would be superfluous to recount the findings of the vast research literature (some of it reviewed by Lyle Bachman, this volume, and in part by other contributors; also see John Carroll, 1983) revealing clear evidence for both a general factor and specific factors in all sorts of mental testing, especially in the measurement of primary and nonprimary language skills. While I do not claim that the pervasive evidence for a general factor (here I have in mind a second order general factor, along the lines of Thurstone's work on intelligence; see John Carroll's [1983] references) conclusively demonstrates the integration of the sort of mental hierarchy (communicative competence) proposed here, this finding is at least consistent with such a proposal, and there are other independent theoretical arguments for such an integration (that is, a unification of the mental hierarchy).

Also, the existence of general factors in the literature on nonprimary language proficiency measurement can probably be interpreted as indicating a second order general factor ultimately linked up with general intelligence, alias communicative competence. In short, the taxonomic models will need to be upgraded into hierarchical modular systems with generative capabilities (see Chomsky, 1982, and Lightfoot, 1982). Incidentally, as I understand Carroll, Vollmer, Farhady, Bachman, Palmer, Upshur, and other contributors to Issues in Language Testing Research (Oller, 1983), I would not expect them to object, in principle, to the sort of hierarchical model proposed here--though we might indeed differ on details concerning how to structure such a model and on the particulars of the present proposal. Also, except for the attempt to make explicit certain relationships within the theory proposed here, it seems to me that the overall approach is entirely compatible with the sort of practical test developments sponsored by Brendan Carroll (1983), Candlin, Leather, and Bruton (1976), and Savignon (1983).

(Originally, by this point, I had expected to draw some invidious comparisons with other theories. However, as noted at the outset, I find myself in agreement for the most part with the practical proposals coming from colleagues at this conference and elsewhere and particularly with the actual testing procedures some of them have recommended [cf. especially Duran et al., 1985]. So it will probably be just as well now to turn to a brief review of the TOEFL and TSE as tests of communicative competence. I will not attempt to repeat what I understand to be the essentially bottom-up approach of Duran et al., but will instead try to reach all the way to the bottom, as it were, with a top-down approach. My comments, therefore, will concern strengths and weaknesses of the present formats as viewed through the filter of the theoretical perspective proposed in this paper.

I will given special consideration to the literacy-oriented portions of the TOEFL, as I was charged to do in the division of labor per Charles Stansfield's correspondence. After that, a few suggestions will be offered concerning a possible programmatic approach to testing new formats and/or improving those now in existence for the TOEFL and TSE.)

VII. The Present TOEFL

In this part, the various sections of TOEFL are reviewed in the order of their appearance, though more attention is given to questions of literacy competence and the literacy-oriented sections. Two specific forms have been examined (Form 3FKTF7 and Form 3EATF12). Before looking at individual sections one by one, I will make a couple of remarks concerning the overall format and the directions throughout.

A. Overall Format

On the positive side, it may be said that the directions are quite clear. A potential problem is that they seem to be aimed at a clientele of literate native speakers at the secondary or postsecondary level. At least some parts look as if they were excerpted, more or less, from established tests written for native speakers of English at the secondary or college level. For instance, consider the following portion of text from the "General Directions":

> Some or all of the passages for this test have been adapted from published material to provide the examinee with significant problems for analysis and evaluation. To make the passages suitable for testing purposes, the style, content, or point of view of the original may have been altered in some cases. The ideas contained in the passages do not necessarily represent the opinion of the TOEFL Policy Council or Educational Testing Service.

All of this material could be omitted from the "General Directions" and be included as a footnote elsewhere, say on the front of the booklet, where the copyright information is presented, and the instructions and examples throughout could be simplified.

Although the syntactic structures of the directions on the whole seem to be at about the same level as the simplest items in the Reading Comprehension and Vocabulary section, the words used and the semantic and pragmatic import of the structures are sometimes more complex than they need to be. For instance, instead of saying,

> You will find that some of the questions are more difficult than others, but you should try to answer every one.

Why not use a simpler structure that says the same thing:

> Some questions are more difficult than others. Try to answer each one.

Instead of,

> You may not go on to the next section and you may not go back to a section you have already worked on.

Why not use imperatives rather than complex declaratives whose comprehension depends on modals and tense markers; e.g.,

> Work on only one section at a time. Do not go back. Wait till you are told to go on.

The whole set of instructions could benefit, I think, from this sort of simplification. Also, greater redundancy could be incorporated through paraphrase. However, it is not helpful, I suppose, to include a wholesale repetition of the material about how to mark answer sheets (more than once) in the middle of the test after everyone should already be doing it. Perhaps they need to be reminded to check for stray marks and to fill in the spaces completely, but if they don't already understand the marking process, wholesale repetition of the previous directions probably won't help.

An obvious weakness in the instructions is the tendency to use words that are more difficult than some of the ones included in the section labeled "Reading Comprehension and Vocabulary." Consider such words as "sample," "supervisor," "advantage," "corresponding," "partial," "oval," "opportunity," "demonstrate," and the like as compared with such words as "aid," "help," "piece," "therefore," "somewhat," "tales," "stories" that appear in the test. If the instruction "Be sure you understand what you are to do before you begin work on a section" (from the "General Directions" on the back cover of the booklet) is taken seriously, an interesting pragmatic bind develops. For examinees at the lower end of the spectrum, the difficulty of understanding the instructions may preclude the possibility of giving correct answers to some of the easier questions. That is, understanding the instructions may require more English than is required for some of the items. The expected effect of this would be to artificially raise the "floor" of the exam in an undesirable way. Although it would affect only a few cases, the possibility could be forestalled, perhaps, by more careful editing of the instructions.

B. Listening Comprehension

A general impression for both forms I examined is that the instructions are given a more natural, phonologically idiomatic reading than the items. It seems to me that the contrast should be reversed.

1. Section 1: Part A

The instructions to the Listening Comprehension section especially could be improved. The fictional context uttered by each of the three speakers at the beginning of the tape serves only to mislead students, since they probably cannot understand the explanation of why the material is included in the first place. Three times in a row, each time from a different speaker, they hear:

> Flight number 53 to Paris will depart from Gate 6 at 9:30 p.m. Will all passengers holding tickets kindly proceed to Gate 6 at this time.

They might well expect (as I did on the first couple of passes) for a sample question of some sort to follow. None follows. There seem to be two purposes. First, to adjust the machine. This could be done with a portion addressed to the proctor and carried out before the examinees ever enter the room, for example, instead of by means of the fictional context that serves no function. Why not:

> The proctor should adjust the volume of the tape recorder at this time.

This instruction could be repeated three times, but I wonder whether this is necessary. The second purpose is to familiarize the examinees with the various voices on the tape. Is this necessary? If so, wouldn't statements such as

> I am one of the speakers on this tape. You will hear my voice several times.

followed by,

> I am another one of the speakers on this tape. You will hear my voice several times.

and so on, deal with the problem in a more comprehensible manner? Although I am a native speaker, my expectation that the thing about the flight to Paris was leading up to a sample item was so strong that I had to go over the tape more than once to figure out why that segment was in there. How this sort of confusion might influence test performance I cannot say, but it probably does not have a desirable effect and should be eliminated if possible.

Apart from the recommendation that the instructions be edited throughout to make them simpler and more comprehensible, the examples of Section 1, Part A, seem more complicated than necessary. Instead of

> Example 1: John is a better student than his brother James.

Why not,

Rewrite 1: John loves to study.

with alternatives (A) John is very tall; (B) John likes fishing; (C) John eats a lot; (D) John likes studying. Then, instead of

Example 2: The truck traffic on this highway is so heavy, I can barely see where I'm going.

something like

Rewrite 2: The rain is so heavy, I can't see.

with alternatives (A) The sun is shining; (B) It's raining hard; (C) The traffic is heavy; (D) It's very foggy. (Shouldn't the weakest students be able to answer the sample questions? Or at least to see that the correct answer is obviously correct and thus to understand what is expected of them?)

Another problem in some of the items is that the correct choice does not always seem quite appropriate. Part of the problem is one of intonation. For instance, question 7 (Form 3FKTF7) begins with the statement, "Franklin Hall was built in memory of Benjamin Franklin," with stress on "Benjamin Franklin." However, the correct alternative, "It was built in honor of Benjamin Franklin" presupposes the topicalization of "Franklin Hall," which has just been upstaged by the primary stress placed on "Benjamin Franklin." This is a subtle pragmatic problem that might be less apt to arise if the reader (the person recording on the tape) were told, right in the script that he or she is reading from, what the correct paraphrase of the statement is.

The same problem comes up in question 7 of Form 3EATF12. The statement is, "They said [contrastive stress here on "said"] they hoped she would soon recover from her operation." Because of the contrastive stress on "said," the speaker seems to doubt what "they said." However, the alternative designated as the correct one, "They wished her a quick recovery" means something entirely different. The best choice, therefore, is not a good paraphrase, given the intonation associated with the stem. (Perhaps we might have a look at the item statistics on this one and compare it with others where the correct alternative is pragmatically equivalent to the stem.)

In question 8 (Form 3EATF12) the speaker says, "Joe would rather wear the blue coat than the grey one," with primary stress on "Joe" and "grey" but even higher contrastive stress on "blue." The correct alternative, "Joe doesn't like the grey coat as much as the blue one," however, is not a necessary implication. The blue coat may be warmer, more suitable for the party, less scratchy, less formal, more formal, and so on. There may be many reasons for preferring it without disliking either coat or liking one better than the other. This time the problem is simply one of finding an appropriate paraphrase that is really implied in a strong way by the stem, rather than simply coming up with something that is marginally acceptable.

In question 10 (Form 3EATF12), "He misunderstood [main stress on "stood"] my instructions," the misunderstanding of the instructions is in focus. However, the keyed alternative, "He doesn't know how to do it" uses anaphoric devices in an inappropriate way. "'It'! What'?" would not be an inappropriate complaint. And, how is it that "his" knowing something can also be brought up as if it had previously been foregrounded (in Wallace Chafe's [1972] sense of the term)? In fact it hasn't been. Here again, I think the intonation invented by the examiner/reader is different from the one intended by the author of the item. Then, in question 12 (Form 3EATF12), "After Carol [primary stress on "Carol"] bought a new dictionary [another primary here], Beth [contrastive stress] decided she needed one too [still higher contrast]." The correct choice is "Beth and Carol needed new dictionaries," but this is not implied by the contrastive stress of "too." It may well be, given this contrastive stress, that Beth didn't need one, but just decided to try to keep up with the Carol.

On the whole the questions in this section present reasonable paraphrase recognition problems. Perhaps they could be improved if the correct choice were made to conform more precisely to the meaning of the stem (or vice versa) as read by the persons on the tape. Or, putting the same suggestion differently, items in general would probably be improved if the facts of the statements could be determined (fixed) more completely. This would help the readers know what intonation to use, and would help those examinees who know more English to get more answers correct. Hence, it should improve test validity, according to the pragmatic theory advocated throughout this paper.

2. Section 1: Part B

In the instructions to Part B, the sample item has an interesting pragmatic ambiguity. Have the students been asked to listen to a certain radio program and discuss that <u>particular</u> program in class, or did the teacher just suggest that they listen to any old program on the radio? In the former case, "Listen to the radio" is not an appropriate response to "What have the students been asked to do?" In any case, it would be good if the facts of the example could be made clearer, and if the example were simplified per the suggestions offered above. The directions that precede the example could also be simplified. By the way, in that connection, can a "voice . . . ask a question"? Or, do we "speak" questions? (To wit: "The question will be spoken . . ." Why not "said"?) These are minor pragmatic problems, but are just the sort of thing the examinee might be gigged for in the second part of Section 2, "Structure and Written Expression." Also, I wonder if the whole effect of the instructions would not be improved if the example were given first. "Now listen to the example. . . . Now select the correct answer. . . . The best choice is (D). . . ." Here all the stuff about how to put the mark on the answer sheet could probably be omitted with no loss.

In form 3FKTF7, question 22 seems to have a peculiar twist due to the man's intonation of "I [primary on "I"] let Barbara borrow [contrastive stress on "borrow"] them." The implication is that "I, not someone else" loaned them out. And because of the higher still contrastive stress on "borrow," there is the implication that the other person thinks Barbara did something else with the notes. However, the correct answer to the question concerns why the man isn't studying his notes to prepare for an exam. This seems strange to me.

In question 27, the conversation is supposedly between a hardware clerk and a customer. Oddly, the customer gives the clerk his entire list of items instead of taking them one at a time, which would, I think, be the more likely procedure. In 34, a man and woman are trying to catch a bus. His shoe is untied. He wants to tie it to keep from tripping. The question is, "What is the man afraid of?" Immediately it may be that he is afraid of falling down (choice A), but the more prominent expectation and the overriding objective is to catch the bus. Hence, if he has an ounce of sense, he's afraid of missing the bus, which is also the concern of the woman, who admonishes him to hurry. Here, it seems to me that the test writer got overly focused on the immediate details of the factual context and forgot that the text implies a broader purpose, an overarching plan to get on that bus!

On the whole, however, this section seems to meet its apparent objective nicely. It provides reasonable listening comprehension and inference problems that are probably at least somewhat characteristic of academic settings. One thing that is not present in all of the exchanges is an element of doubt, or conflict, or disequilibrium that would serve to focus the listener's attention and help to motivate the interaction of the two interlocutors.

3. Section 1: Part C

It seems to me that these extended talks have the most potential for authenticity and yet fail to realize it for lack of imagination on the part of item writers. Whoever does these things should have some ability to conceptualize exchanges that might really occur--interactions with some meaning in them. The sample lecture, a talk about Ellis Island, is too long and deadly dull. Who cares about Ellis Island? Or how people got there? It's simple enough, probably, but could be improved by shortening and by inserting something (almost anything) interesting--for example, having a Boy Scout drown nearby or infesting it with rare deadly snakes or having it erupt into a volcano. Or take a different island in the South Pacific and maroon George and Mary there. Actually, if we are going to deal with the academic setting, the real items should pertain to suitable contexts at colleges or universities. But even these can be interesting.

Telling students in advance which questions go with which minitalk could be avoided by labeling them "Talk A," "Talk B," and so on and by indicating break points in the test book between items. The present instructions are distracting, especially on the first pass.

On Form 3FKTF7, the first talk (questions 36-43) is strange. Why does the man care whether or not the woman has signed up for or paid for a dorm room yet? Why is he so concerned about her finances? The whole dialogue seems unmotivated and unmotivating. It isn't the kind of thing we normally take notes on, but if the student isn't allowed to take notes, it will be difficult to remember some of the useless details that will later appear in test questions. The talk about Alaska is a little more plausible, but no less dull. The details seem to be thrown in at random, as if written by someone who knows little about Alaska and cares less, but who knows how to use an encyclopedia or at least to talk like one. The switch to "environmental science" is implausible. Who cares what the speaker was doing in Alaska twenty years ago? On the other hand, some of the questions can be answered without having paid attention to the lecture--for example, the one about what Anchorage is like today, or about the pipeline and the housing shortage.

Two of the talks in Form 3EATF12 seemed better than the ones in Form 3FKTF7. The talk about writing a paper seemed more realistic than the others in spite of the fact that the questions concentrated on details that were not of any general interest or value--such as how many sources should you consult in your library work? This is really an unanswerable question in general, but the lecturer does specify four sources. The talk on the whole seems realistic and has some authenticity. The question about when the lecture in question is taking place is marginally significant because the assignment is to be done over a break. However, I think most examinees could guess the answer to the question, "Why doesn't the teacher want handwritten papers to be turned in?" without having heard the lecture. One minor drawback is that it sounds like it is being read, and I doubt that it would be read in a real-life setting.

The lecture on glue is good. I liked this one because it departed from the sort of banal trivia that occupied the lecture about Alaska. Much of the information in this talk was either new to me or expanded on things I knew about. Therefore, it seemed to me to have some merit as an attempt to communicate "by bridging an information gap" in the sense of Harrison (1983).

The third talk in Part C of Form 3EATF12, however, returns to a topic that is overworked, tired, and boring. It lacks any flare of imagination, it does not cross any significant gaps, and I might have dozed off if I hadn't been taking notes. As it was, I had to wake up my pen. I thought it had run out of ink, but it had just fallen asleep!

C. Structure and Written Expression

As in the case of the previous section, here too, it seems that the example in the instructions could profitably be made simpler. Why write,

Mt. Hood ------ in the state of Oregon.

rather than something simpler such as

George ------ tall.

with alternatives such as (A) are, (B) is, (C) has, (D) can. Here the correct choice would be quite obvious even to the weakest candidates. In the other case, it is less transparent.

On the whole, the questions that are concerned primarily with the sequence of elements in the first part seem okay though uneven in some cases in difficulty and focus and especially uneven when compared with some of the items in the editing format. Compare item 5 (an ordering item that seems to hinge on stylistic preferences) with 12 (both drawn from the sequence-oriented section of Form 3ATF12):

5. Although believed by most Americans to be a desert, ------.

(A) the state of Utah has one-third of its area covered by deep, lush forests [this choice seems okay to me except that it is less elegant than the next one]

(B) deep, lush forests cover one-third of the area of the state of Utah

(C) covering one-third of the state of Utah are deep, lush forests

(D) but deep, lush forests covered one-third of the state of Utah

That one seems fairly subtle and perhaps might be missed by some natives, but consider 12:

12. Kenyon Cox, an art critic and painter, ------ best known for his murals, portraits, and decorative designs in public buildings throughout the United States.

(A) who is

(B) he is

(C) is

(D) and he is

This item is dependent mainly on obligatory surface syntax while 5 depends on a stylistic judgment. Moreover, as Duran et al. (1985) observe, quite a few of the items in the editing format also depend on relatively simple surface constraints--such as number agreement between subject and verb.

Four years ago, Clifford and Lowe (1980) put forward the very plausible hypothesis (which they supported with careful research on the FSI oral proficiency interview) that low-level operations of surface syntax and morphology tend to be mastered early (that is, reach a ceiling and level off), while other aspects of language proficiency, such as vocabulary knowledge, style, and the like continue to develop. Perhaps this is why it is apparently necessary to beef up the editing part of the Structure and Written Expression section with more subtle questions concerning style. For example, consider item 26 from Form 3FKTF7:

26. $\underset{\text{A}}{\underline{\text{After the revolutionary War}}}$, Mount Vernon $\underset{\text{B}}{\underline{\text{was}}}$

$\underset{\text{C}}{\underline{\text{enlarged and made bigger}}}$ $\underset{\text{D}}{\underline{\text{by}}}$ George Washington.

Although this sort of item may achieve the desired psychometric effect, one wonders about using such disparate techniques in the same subsection, "Structure and Written Expression." A response set develops as one works through the items, and this one seems out of place. One does a double take. Would it be possible to distinguish the type of variance generated by items such as this one and that generated by items aimed at surface syntax? The very useful factor analyses by Swinton and Powers (1980) might help to answer this question, but it is difficult to be certain without access to the actual test items (from a TOEFL administration in November 1976, Form YTF4) that formed the basis for their research and that unfortunately are not given in their paper. Although their research differentiated seven widely disparate language groups (African, Arabic, Chinese, Farsi, Germanic, Japanese, and Spanish) and used a form of TOEFL with five subsections (Listening Comprehension, English Structure, Written Expression, Vocabulary, and Reading Comprehension), the factorial structure, which seemed to fit all groups well except for the Farsi speakers, showed three distinct orthogonal factors.

The first was clearly a Listening Comprehension factor for all seven language groups. The second factor was somewhat ambiguously defined by items from the Written Expression, Reading Comprehension, and Structure sections, and the third factor for most language groups was best defined by Vocabulary items (and sometimes Reading Comprehension items as well).

Although this three-factor breakdown may have been used properly to help justify the revised three-section format of TOEFL, one might like if a finer distinction between items aimed at various aspects of proficiency, (e.g., inflectional morphology; lexical knowledge, e.g., items aimed at derivational morphology, breadth of vocabulary), and items aimed at within-clause versus across-clause syntactic constraints and semantic, pragmatic, and stylistic considerations of various sorts. In such a case, perhaps it would be possible to define a more delicately articulated basis for classifying items. At any rate, no other testing agency dealing with nonprimary language proficiency measurement is in as good a position to carry out such research as are the ETS people working with the TOEFL.

Whereas the present distribution of items in the "Structure and Written Expression" section makes sense as a practical grouping, there are theoretical reasons to suppose that elements within it might be profitably differentiated for many purposes, but especially for diagnostic ones. In a sense, this section is the one that as linguists perhaps we should know the most about, and yet, surprisingly, it may be the most mysterious section of all. Perhaps the taxonomy proposed by Duran et al. (1985) will afford some new insights into the various items included in the Structure and Written Expression section.

D. Reading Comprehension and Vocabulary

This part of the paper deals with the third and final section of the TOEFL (the section that I was asked to concentrate on in the correspondence from Charles Stansfield). It is subdivided into two subparts. In the first part, the vocabulary items are discussed and, in the second, items based on written passages are reviewed. I noticed that one of the two forms I examined, Form 3FKTF7, also had a few items in a third format. Instead of a passage followed by the traditional questions, the stem consisted of a single sentence, and the alternatives consisted of possible paraphrases for that sentence. This is an item type we used at UCLA some years ago. It consistently gave reliable and useful information, but Stansfield informs me that it has been deleted from the TOEFL specifications for the "Reading Comprehension and Vocabulary" section, though it is widely used in the "Listening Comprehension" section. It is no longer used in "Reading Comprehension and Vocabulary," for instance in Form 3EATF12, or in the form reviewed by Duran et al. (1985), Form 3FATF5.

1. Vocabulary

As in the case of previous sections, the example for the vocabulary items could profitably be simplified for the same reasons given earlier. Is "ordinary"--a word that is the focus of the sample item--that much less difficult than "somewhat" and "therefore"--words that appear in actual items in Form 3FKTF7? Or what about "favorable" and "plane," which are used in items in Form 3EATF12? The more transparent the example, the more obvious the correct answer is to the examinee, and the more successful will be the communication of instructions concerning what needs to be done.

In general, (following directly from the theory outlined above) vocabulary items embedded in context are bound to be better than items not so embedded and in general (up to some unknown point of diminishing returns), the more context the better.[3] The reason for this is that additional context helps to achieve the purpose of "fixing the facts." However, it would seem wise to do some careful research on this question and to examine the factorial distribution of items embedded in progressively greater amounts of context versus those in progressively less context up to the limit of so-called isolated vocabulary items. I have argued elsewhere (especially in my response to Farhady, 1983) that there is no such thing as a truly "discrete-point" vocabulary item. I won't reiterate that case here, but it may be worthwhile to note in passing that isolated vocabulary items may have some significant merits and should not necessarily (at least, not at this stage) be categorically rejected. My own expectation, however, is that items embedded in useful, interesting, authentic academic contexts will yield better results on the whole. They will be easier to write, less ambiguous, and overall more valid by both norm-referencing and criterion-referencing requirements. Greater construct validity will produce net advantages all around.

However, the items cannot be any better than the texts in which they are embedded. To the extent that the texts are boring, listless, and unmotivated (apropos of the theory presented in Oller & Richard-Amato, 1983, Chapters 1 and 2 especially), it is to be expected that the items will also be lackluster. The matter of "fixing the facts" comes into play here again as everywhere else. If the text is uncertain in its relation to the facts, the item may also be wobbly as to which alternative is the best choice. For instance, consider a couple of items from Form 3FKTF7:

2. A simple rheostat consists of a wire wound around a cylindrical piece of insulating material.

   (A) portion

   (B) coil

   (C) cell

   (D) fiber

On this one, I ran into some difficulty because I know that the thing with the wire wrapped around it is in fact called the "coil," so this item is pragmatically attractive. True, the word "portion" may be the technically preferred alternative, but is it reasonable to say that a person who chooses "coil" knows less English than the one who chooses "portion"? Actually, I don't think that "portion" is a good synonym for "piece" when the object is a discrete (countable) entity as opposed to a continuous (uncountable) substance. On the other hand, if there were some additional text where the small cylindrical piece of material were differentiated from the coil, which consists of it plus the wire, the problem would be eliminated. More text will tend to reduce ambiguity.

Another example can be found in item 5 in the same form:

5. Chlorpromazine is a drug prescribed by doctors to reduce agitation.

   (A) swelling

   (B) nervousness

   (C) infection

   (D) discomfort

Here again, I ran into difficulty partly because of what I didn't know (namely, what "chlorpromazine" was) and, because of the lack of any disambiguating context, I had a harder time choosing between the alternatives. I was especially attracted to (D) and lingered momentarily over (A). I figured that "swelling" in a wound or sore might be reduced by a drug, and the terms "infection" and "discomfort" lent some support to the possibility that what was at stake was some sort of localized injury. I was, for some reason, inclined to read "agitation" (the stem word) as meaning (A), (B), or (C). Although I know its nonconcrete, mental meaning very well, the distractors, together with the limited context and my lack of knowledge of what chlorpromazine was, led me to try first to work with the hypothesis that the drug was used in the case of physical injuries. Now, if I had encountered the alternatives in isolation, I don't know whether this would have happened. In fact, I think I caught my error and changed my choice to (B) after checking the stem against the alternatives without rereading the context. Having done this, I further narrowed down the use of the drug. Though I had not yet looked it up, I was quite certain it is used in the treatment of mental disorders. (Maybe I had even heard that before; I am not sure.) At this point, I stopped to look it up. Ouch! It wasn't in my Webster's New World College Dictionary. Does this tell us anything about the usefulness of the item? A little more context would have resolved the trouble.

For all of the foregoing reasons, it would make sense to experiment with vocabulary items embedded in cloze-type contexts. I cannot imagine what could be lost by such experimentation, and much might be gained. Nonetheless, it is important to remember the general caveat that the test can be no better than the text on which it is based. Of course, if items in isolation turned out to work better than items embedded in cloze-type contexts of increasing lengths (other factors, such as topic, difficulty, and the like, being controlled, notwithstanding the strange conclusions of Porter [1983], who deliberately allows all these other variables to wander aimlessly while he claims to assess the effects of the variable "length of text"), the theory I advocate here and elsewhere would have to be abandoned or radically revised. However, contrary to some unusual interpretations of the research literature by Alderson (1983, and elsewhere) and the theory-less wanderings by Porter (1983), all the available evidence supports the claim that increasing the amount of text beyond the traditional five-to-ten word limit on either side of a blank (other things being equal) makes it easier to guess the correct answers on a cloze test. Further, the sort of pragmatic theory advocated here suggests why, other things being equal, additional context will also increase test validity. On this point, I do not think, as Alderson (1983) claims, that the research is ambiguous (also see Chavez-Oller, Chihara, Weaver, & Oller, 1985).

2. Reading Comprehension

We come now to the section of TOEFL where passages of text are used (except for three items--58, 59, 60--at the end of Form 3FKTF7) and where the focus of items is more or less on the content of passages rather than on individual words. Because of the greater expanse of material (text), we might expect the facts to be better fixed (as discussed above in the theory proposed in this paper), and we might expect these items to be among the most effective on the entire test.

Passing quickly over the standard observation that the sample item--in this case, about the gorilla--is too complex (unnecessarily long, uses more difficult words than necessary), let us turn immediately to an examination of the items in Form 3FKTF7 and Form 3EATF12. As others have observed (especially Duran et al., 1985), there is a tendency for items here and elsewhere in the test to have an encyclopedic flavor. The passages are on assorted topics, expository in character, ranging from important scientific information (for example, how adrenaline is released in the blood and what effects it has) to trivia (for example, superstitions about gem stones and how they are still reflected in certain practices today). Most of the choices are dry and somewhat academic (for example, the selection from a university catalog about how to apply to business school) while others seem almost touristy (for example, the piece about the National Gallery of Art), bringing to mind the musty odor of museums and sleepy afternoons at the zoo.

What is generally missing in most of the passages is any element of disequilibrium, doubt, puzzlement, surprise, or conflict to motivate the text itself (that is, its having been written in the first place) or the reading of it. Some of the texts do not seem authentic. Like the lecture in the Listening Comprehension section about Alaska, they seem to wander aimlessly instead of focusing on some significant purpose. If item writers could be instructed to look for texts that focus on conflict, in the sense of Peirce, Dewey, and the theory discussed above in part III (also see Oller & Richard-Amato, 1983, Chapter 1), I think the whole quality of the TOEFL could be upgraded and that it would be more authentic, relevant, engaging, challenging, and, in the end, more valid.

The problem of motivation and interest deflates the usefulness of some items. For instance, in item 32 of Form 3EATF12, who really cares whether it was in the fourteenth century, the sixteenth, the eighteenth, or the twentieth when "modern beliefs about birthstones are thought to have originated." "Thought to have originated" indeed. By whom? On the basis of what evidence? No contrary views? The fact is that it doesn't seem to make any difference to anyone. Nothing hangs in the balance. There is no goal or objective that will be lost, no person who stands to lose or gain, no significant conflict to be resolved by this particular fact. It is an incidental piece of trivia. Also, one must be careful in answering because the text says "1700s" while the alternative says "eighteenth century." In fact, certain other items require even more complex mental computations than this one; for example, item 58 on Form 3EATF12, which requires subtracting 1700 from 1980 to get 280 years, or 59, which requires adding "seventeen thousand" to "five hundred" in order to obtain "17,500"; though, again, there is no motivation (that is, a conflict-based need, in the sense of the theory of communicative competence adopted above) for knowing any of these values. As a result, items of this sort seem to lack authenticity.

In addition to the problem of achieving authenticity in the specific items cited, a more general problem is the quality of the texts themselves. The theory advocated here suggests that any text will have some authenticity for some purpose, but that the degree to which it is apt to capture the interest of examinees and thus to engage and challenge their communicative competence will be enhanced to the extent that the text contains some more or less central element of disequilibrium or conflict. This conflict factor will help to focus the attention of both test takers and test item writers on the important facts rather than on trivial details. Also, this attention-focusing mechanism will make the facts themselves more vivid and noteworthy and thus will help to fix the facts in the required sense.

In addition to the general problem of text selection, there is the ever-present difficulty of making the keyed choices really the best, fitting them to the facts as specified (fixed) by the texts. If the texts are relatively unfocused, even rambling--for example a professor holding forth to no apparent purpose on Alaska (Form 3FKTF7), the bit of trivia about birthstones (Form 3EATF12), the piece on Freud's theory of negation (Form 3FKTF7), or the paragraph about commerce and communication during the American colonial period (Form 3FKTF7)--not only will it be more difficult to say what the relevant facts are, it will per force be more difficult to form meaningful questions about those facts. For example, can it really be said with any confidence that "In the passage [the one about Freud's theory of negation], the phrase 'rediscovered in perception' is used in order to show that in (C) the term 'reality' is based on human perceptions." This is the keyed answer to item 54. However, the text does not deal with the "term 'reality'" [my emphasis] at all, but with the concept of reality and how it is acquired. More specifically, the text concerns Freud's quandary about how one distinguishes not the term "reality" from other terms, but what is real from what is not. Any epistemologist will recognize that there is a difference of importance here. However, the important issues in the text for questions 51-57 are not very obvious because the passage (as excerpted from the original) leaves the critical issues (the doubtful ones) largely undefined and out of focus. In fact, the concluding half of the passage asks: "When does the ego acquire the ability to negate? Is saying 'no' connected to the acquisition of language? Freud did not stress that the ability to negate is directly connected to the development of the self. . ." (p. 21 of Form 3FKTF7). Instead of clarifying, these remarks wander off in several different directions drawing attention away from the question Freud was actually grappling with.

VIII. The TSE

In this brief section, the Test of Spoken English (TSE) is examined. The purpose is to make recommendations for the improvement of this instrument. Let it be noted first of all that the sample tape reveals a sound basis for evaluating general proficiency in spoken English--which seems to have been the goal of the test from its inception. However, the test is not without certain potential weaknesses, if viewed from the vantage point of the theory advocated in this paper (or perhaps even from a Hymesian taxonomical viewpoint).

The strongest parts of the test would seem to be Sections 4 and 5. Section 4, the part based on a series of visually depicted events, has the advantage of a sequentially developing, temporal context. This provides a great deal of fact-fixing information that helps to make the task meaningful and relatively easy to grade. Section 5, consisting of a single picture displaying a minor disaster (at least in the Examinee Handbook [1984]), has the added feature of motivation. Since the picture displays a situation that could be factual, this part of the test also benefits greatly from the advantages of determinate facts--and an implicit series of events culminating in an undesirable outcome. The fact that the events lead to such an end (the driver smashing the kid's bike with the car) motivate talk about this situation in the required sense of the theory advocated in this paper.

In addition to these two parts, Section 2 also has certain merits. It portrays temporal development, and the sample text in the Examinee Handbook contains a certain element of disequilibrium (mild conflict) that serves to motivate the text. What might be criticized is the fact that this section does not involve "real" (dyadic) communication. However this portion has merit and helps to lower the "floor" of the examination in a desirable way.

However, the other sections of the test (excluding Section 1, which is not scored, though perhaps it should be) are less obviously well structured or well motivated. Section 3, which jumps from topic to topic arbitrarily, lacks the sort of temporal development that is characteristic of normal experience and ordinary communication. It may work very well as is, but I think that at least some students would render a considerably different performance on a task that was more motivated.

Section 6, if anything, is even less well structured and motivated. "Describe a telephone in detail." Hmmmm. What a strange request. I suppose I could do it, but it is hardly the sort of thing that one might be asked to do in normal communication. It is an extraordinary task. In fact, it reminds me of those horrible essay tasks where you sit and stare at a painting until nearly nodding off and then you give up and hunt up another one, hoping that it will elicit a spontaneous flow of verbal creativity. Rarely does such a thing produce any spark of brilliance, however. The suggestion, "Describe the things that make a perfect day" seems more interesting, but it too lacks the disequilibrium necessary to talk about over coffee. "What is your opinion of the problem of automobile pollution?" On this one, most of us can be brief. It's too bad, but we can't seem to figure any way around it.

Wouldn't some more conflict-laden questions, or leads, generate better (more respresentative) performances? Instead of the bit about a telephone, why not, "Tell about a person you can't stand to be around." Or, "Tell about someone you have a hard time getting along with." Or, "Tell about someone who doesn't like you." Instead of saying what makes a perfect day, "Tell about the worst day you ever had." Or, "Tell about the most difficult trip you ever made." Instead of asking for an opinion about automobile pollution, why not ask, "How do you feel about the fact that automobile pollution is a major cause of lung cancer?"

Section 7 runs into the same sort of trouble. It is dull and unmotivating. The task could be brought down to where the proverbial rubber squeals on asphalt by having the student tell about something that is a little more engaging on a personal level. Something along the lines of a conflict-laden sociodrama (see Part IV of Oller & Richard-Amato, 1983) is called for rather than reciting a list of details from a course schedule. It's hard to get excited about the latter. On the other hand, suppose we take a lead-in that has just a little conflict in it: "You are a creative writing major. You want to take Mr. Smith's Creative Writing 201 that meets at 3:00 to 4:00 p.m. on Monday, Wednesday, and Friday. However, your counselor has advised you to sign up for History 101, which meets at the same time on the same days. He is unaware of the conflict. Explain how you would handle the problem and why you would do it that way."

Innumerable scenarios that contain some element of motivating disequilibrium could easily be thought up, though it is apparent that the importance of "conflict" as a motivator of communication may have only dimly occurred to TSE item writers. Other formats that might be explored for possible use in the TSE would include text-based varieties of imitation, paraphrase, interpolation, extrapolation, retelling, and the like. For more concrete suggestions along this line, see Oller (1979).

Concluding Remarks

Up to this point, I have concentrated on what seem to be weaknesses only because I think they may be useful in making a couple of good tests even better. Also, discussing such weaknesses may help in clarifying the practical implications of the theory I have been advocating throughout. However, as it now stands, the TOEFL and TSE are obviously doing a fairly good job of assessing some aspects of communicative competence. Among those are skills in speaking, listening, and reading. It is not clear to what extent the TOEFL may be an indicator of productive skills (speaking and writing), and though the TSE does measure speaking skills to some extent, writing skills are only assessed indirectly by the TOEFL. It would seem that some TWE (Test of Written English) parallel to the TSE might be advised. With reference to the literacy-oriented sections of TOEFL, particularly, the Reading Comprehension and Vocabulary section, it would seem that the present format is adequate for a wide range of purposes, but probably not for measuring essay writing skills.

My recommendation for the literacy-oriented portions of TOEFL in particular is to investigate a wide range of item types, including items with varying degrees of text-embeddedness aimed at lexical knowledge, including comprehension of syntactic units smaller than words (i.e., inflectional morphemes and derivational morphemes) and larger than words (idioms, collocations, routine phrases, and larger constructions), items aimed at the utilization of grammatical (syntactic, semantic, and pragmatic) constraints within the across-clause boundaries, as well as items aimed at all of the conceivable elements of the Duran et al. (1985) taxonomy of evaluative criteria.

Also, I recommend carefully planned studies of the psychometric properties of texts with varying degrees of coherence and authenticity as judged by competent evaluators operating on the basis of clear theoretical criteria. Specific hypotheses concerning degree of conflict focus, for example, can be tested--e.g., the more conflict-focused the text and the more determinate the facts, the better the items (other things being equal). Procedures for generating item types, or potential item types, such as the vast array of possibilities under the general category of cloze procedure, and also paraphrase, seem especially promising for use in the literacy-oriented sections of the TOEFL (specifically, "Structure and Written Expression" and "Reading Comprehension and Vocabulary"). At least cloze items ought not to be ruled out without careful examination. In addition, there are other general techniques that may be applied to problems of inference presupposition, association, and implication--to assess an ability to "read between the lines."

As it now stands, judging from the careful evaluation of Duran et al. (1985) and from other studies of the TOEFL (especially the work of Swinton and Powers, 1980, and, before them, the research by Pike, 1979--see also Hale, Stansfield, and Duran, 1984, for summaries of a great deal of additional work), everything points to the conclusion that the TOEFL is presently a fairly good measure of communicative competence and that it is augmented in a desirable way by the TSE. Beyond this it is to be expected that the ongoing research promises to make these tests increasingly better. I concur with Canale (1983), Cummins (1983), and Krashen (1982) that the major research problem for the immediate future is to develop and refine a theoretical perspective that may then serve as a basis for planning and interpreting experimentation.

Notes

[1] This paper was prepared for the Second TOEFL Invitational Conference. Although it represents a theoretical bent that is no doubt peculiar to the author, it has nonetheless been influenced by many sources. Some of those need to be acknowledged here.

The most central sources are to be found in the writings of the American pragmatists, notably, C. S. Peirce and his proteges, William James and John Dewey. Of the latter, Dewey's work has been more influential than that of James, but neither exceeds the influence of their mentor, Peirce. Related to their writings and helping to set apart the theoretical position they represent is the work of Bertrand Russell, whose epistemology is considerably more skeptical than that of the Americans concerning the accessibility of the real world to human understanding. Einstein's popular writings on reality, experience, language, and communication have also been influential though he wrote a great deal less than the others. The course of de Saussure on general linguistics had a lesser impact, though an important one, and work from more recent decades by Jean Piaget and Noam Chomsky have had a profound impact, especially Chomsky's ideas concerning the biological aspects of language acquisition as opposed to those of Piaget, as well as Chomsky's insistence on a generative framework (see Piatelli-Palmarini, 1980). A general perspective on these sources may be found in (Oller, 1986)--an edited book of readings containing work by all of the above mentioned authors along with explanations comment.

Although the pragmatic theory, which I advocate here and elsewhere, is not unaffected by the parallel development of Hymes' ideas--especially as interpreted and expanded by such second language researchers as Swain, Canale, Cummins, Bachman, Palmer, and Upshur--this pragmatic theory stems from a distinct epistemological basis. Nevertheless, I am very much indebted to Duran, Canale, Penfield, Stansfield, and Liskin-Gasparro (1985) for their monumental study of the TOEFL and especially to four graduate students--Fred Davidson, Brenda Hayashi, Thom Hudson, and Bryan Lynch--for comments on this theoretical perspective during an advanced seminar in language testing at UCLA in the spring quarter of 1984.

[2] Frankly, I was glad to see that this idea is apparently agreed with by Duran, Canale, Penfield, Stansfield, and Liskin-Gasparro (1985). See their discussion on page 24, where they say, ". . . the range and complexity of skills required on various sections of the TOEFL was directly related to the amount of language and to the semantic and textual complexity of TOEFL items. The more language used, and the more authentic this language, the greater the number and kinds of communicative skills. . ." (p. 24). Of course, the authors' use of the phrase "number and kinds of communicative skills" is indicative of the bottom-up, taxonomic approach. It is also important, I think, that they note, "At each stage in the process of designing successive versions of the skills checklist, the inadequacies of a taxonomic approach to communication became more

apparent" (p. 12). In other words, the authors clearly recognize the need for top-down reasoning as well and ultimately for a coherent theory of communication and communicative competence.

[3] I understand that Duran, Canale, Penfield, Stansfield, and Liskin-Gasparro (1985) lean in this direction too. See note 2 above.

References

Alderson, J. C. (1983). The cloze procedure and proficiency in English as a foreign language. In J. Oller (Ed.), Issues in language testing research (pp. 205-217). Rowley, MA: Newbury House.

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.

Candlin, C. (1981). Discoursal patterning and the equalizing of interpretive opportunity. In L. Smith (Ed.), English for cross-cultural communication (pp. 151-165). New York: Macmillan.

Candlin, C., Leather, J. H., & Bruton, C. J. (1976). Doctors in casualty: Applying communicative competence to components of specialist course design. International Review of Applied Linguistics, 14, 245-272.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. Oller (Ed.), Issues in language testing research (pp. 80-107). Rowley, MA: Newbury House.

Carroll, B. J. (1983). Issues in the testing of language for specific purposes. In A. Hughes & D. Porter (Eds.), Current developments in language testing (pp. 109-114). London: Academic Press.

Chafe, W. (1972). Discourse structure and human knowledge. In J. D. Carroll and R. O. Freedle (Eds.), Language comprehension and the acquisition of knowledge (pp. 41-69). Washington, DC: V. H. Winston.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W., Jr. (1985). When are cloze items sensitive to constraints across sentences? Language Learning, 35, 181-206.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: Massachusetts Institute of Technology.

Chomsky, N. (1975). Reflections on language. New York: Pantheon.

Chomsky, N. (1980). Rules and representations. New York: Columbia University.

Chomsky, N. (1982). Some concepts and consequences of the theory of government and binding. Cambridge, MA: Massachusetts Institute of Technology.

Clifford, R., & Lowe, P.. (1980, May). Language proficiency testing: A dynamic model versus a static model. Presented in two parts at the Second International Language Testing Symposium, Darmstadt, Germany.

Cummins, J. (1983). Language proficiency and academic achievement. In J. Oller (Ed.), Issues in language testing research (pp. 108-126). Rowley, MA: Newbury House.

Curtiss, S. (1977). Genie: a psycholinguistic study of a modern-day "wild-child". New York: Academic Press.

Dewey, J. (1916). Essays in experimental logic. New York: Dover.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. E. (1985). The TOEFL from a communicative viewpoint on language proficiency: A working paper (TOEFL Research Report 17). Princeton, NJ: Educational Testing Service.

Einstein, A. (1954a). Physics and reality (1936). In Out of my later years (pp. 53-97). New York: Philosophical Library.

Einstein, A. (1954b). The common language of science (1941). In Out of my later years (pp. 111-113). New York: Philosophical Library.

Farhady, H. (1983). The disjunctive fallacy between discrete-point and integrative tests. In J. Oller (Ed.), Issues in language testing research (pp. 253-269). Rowley, MA: Newbury House.

Fodor, J. A. (1975). The language of thought. New York: Crowell.

Fodor, J. A. (1980). Representations. Cambridge, MA: Massachusetts Institute of Technology.

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). Summaries of studies involving the Test of English as a Foreign Language, 1963-1982 (TOEFL Research Report 16). Princeton, NJ: Educational Testing Service.

Haley, J. (1963). Strategies of psychotherapy. New York: Grune and Stratton.

Harrison, A. (1983). Communicative testing: jam tomorrow? [sic] In A. Hughes & D. Porter (Eds.), Current developments in language testing (pp. 77-86). London: Academic Press.

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

James, W. (1907). Pragmatism. London: Longmans.

Jones, L. 1979. Functions of English. In C. J. Brumfit and K. Johnson (Eds.), The communicative approach to language teaching (pp. 223-226). Oxford: Oxford University Press.

Krashen, S. (1982). Principles and practice in second language acquisition and teaching. Oxford: Pergamon.

Lightfoot, D. (1982). The language lottery: toward a biology of grammars. Cambridge, MA: Massachusetts Institute of Technology.

Morrow, K. E. (1977). Techniques of evaluation for a notional syllabus. Reading, England: University of Reading, Center for Applied Language Studies (mimeo).

Oller, J. W., Sr. (1963). La familia Fernandez. Chicago: Encyclopedia Britannica Educational Corporation.

Oller, J. W., Sr. (1965). Emilio en Espana: el espanol por el mundo. Chicago: Encyclopedia Britannica Educational Corporation.

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

Oller, J. W., Jr. (Ed.). (1983). Issues in language testing research. Rowley, MA: Newbury House.

Oller, J. W., Jr. (1986). Language and experience: A book of readings. Albuquerque: University of New Mexico.

Oller, J. W., Jr., & Richard-Amato, P. (Eds.). (1983). Methods that work: a smorgasbord of ideas for language teachers. Rowley, MA: Newbury House.

Peirce, C. S. (1877). The fixation of belief. Originally in Popular Science Monthly, 12, 1-15. In C. Hartshore and P. Weiss (Eds.), Collected papers of C. S. Peirce: Volume V pragmatism and pragmaticism (pp. 223-247). Cambridge, MA: Belknap Harvard University.

Peirce, C. S. (1878). How to make our ideas clear. Originally in Popular Science Monthly, 12, 286-392. In C. Hartshore and P. Weiss (Eds.), Collected papers of C. S. Peirce: Volume V pragmatism and pragmaticism (pp. 248-271). Cambridge, MA: Belknap Harvard University.

Peirce, C. S. (1905). Issue of pragmaticism. Originally in The Monist, 15, 481-499. In C. Hartshore and P. Weiss (Eds.), Collected papers of C. S. Peirce: Volume V pragmatism and pragmaticism (pp. 293-313). Cambridge, MA: Belknap Harvard University.

Piaget, J. (1947). Social factors in intellectual development. In The psychology of intelligence (pp. 156-166). Tottowa, NJ: Littlefield Adams.

Pike, L. (1979). An evaluation of alternative item formats for testing English as a foreign language (TOEFL Research Report No. 2; ETS Research Report No. 79-6.) Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 206 627.)

Popham, W. J. (1975). Educational evaluation. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1981). Modern educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Porter, D. (1983). The effect of quantity of context on the ability to make linguistic predictions: A flaw in a measure of general proficiency. In A. Hughes & D. Porter (Eds.), Current developments in language testing (pp. 63-74). London: Academic Press.

Rivera, C. (Ed.). (1983). An ethnographic/sociolinguist approach to language proficiency assessment: Multilingual matters 8. Avon, England: Multilingual Matters.

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

Swinton, S. S., and Powers, D. E. (1980). Factor analysis of the Test of English as a Foreign Language for several language groups. Princeton, NJ: Educational Testing Service.

Test of Spoken English. (1984). Examinee handbook. Princeton, NJ: Educational Testing Service.

Watzlawick, P., Beavin, J., and Jackson, D. (1967). Pragmatics of human communication. New York: Norton.

COMMUNICATIVE COMPETENCE AND TESTS OF ORAL SKILLS

Dan Douglas

INTRODUCTION

My purpose in this paper, which is in the general category of suggestions for the improvement of current measures and the development of new measures in a communicative competence framework, is to discuss the TOEFL Listening Comprehension section and the Test of Spoken English as measures of oral skills. I will try to suggest ways in which these instruments might be improved as measures of communicative competence by discussing work being carried out at the University of Michigan English Language Institute on the development of a new listening comprehension test and analysis of performances on part of the TSE/SPEAK at the Wayne State University English Language Institute. I will make these suggestions in light of a background discussion derived from interlanguage studies in general and from research on domain-specific interlanguage use in particular, elaborating a concept of "discourse domain" (Selinker and Douglas, 1985) that has consequences for our attmepts to construct measures of communicative competence.

I take as a starting point the communicative competence framework of Canale and Swain (1980) elaborated on by Canale (1983), though I will try to focus a bit more than they do on reception in communication in my discussion of the Listening Comprehension section of the TOEFL. It seems to me to be particularly important to focus on learners' problems with perception in the target language since there has been an unfortunate lacuna in work on communicative competence regarding receptive skills. Certainly, very little is known about the listening process, the problems learners have with comprehension in the target language, and the methods they employ in overcoming these problems. This, at a time when the notion of comprehensible input (Krashen 1978; Terrell 1980) is of such concern, seems to be a particularly serious gap in our understanding of communicative competence. However, the decision by Duran et al. (1985) to omit consideration of strategic competence in their analysis of the TOEFL was, in my view, unnecessary. If we are interested in analyzing a test within a communicative competence framework, then any item that requires the subject to demonstrate comprehension, whether native-like or not, should be susceptible to analysis. For example, the table of examples of communication strategies in Tarone, Cohen, and Dumas (1983) will look very familiar to writers of test items: Phonological, morphological, syntactic, and lexical variations on such strategies as transfer, paraphrase, and reference to prefabricated patterns are often the basis for distractors in multiple-choice item writing.

Figure 1

Communication Strategies of Second Language Learners
(Tarone, Cohen, and Dumas, 1983)

From "A closer look at some interlanguage terminology: A framework for communication strategies." In C. Faerch and G. Kasper, (Eds.), Strategies in interlanguage communication (pp. 4-14). New York: Longman.

	Phonological	Morphological	Syntactic	Lexical
Transfer from NL	*/ʃip/* for */ʃɪp/*	*The* BOOK OF JACK for *Jack's book*	*Dió* A ELLOS for LES *dió* A ELLOS In Spanish-L2	*Je* SAIS *Jean* for *Je* CONNAIS *Jean* in French-L2
Overgeneralization	*El carro*/karo/ *es caro* (Flap r generalized to trill contexts — Span.-L2)	*He* GOED *Il* A *tombé* in French-L2	*I don't know* WHAT IS IT	*He is* PRETTY (Unaware of the semantic limitations)
Prefabricated pattern	—	—	*I don't know how do you do that*	—
Overelaboration	/hwʌt ar ju duɪŋ/ for /wʌtʃəduɪn/	*I* WOULD NOT HAVE *GONE*	YO *quiero ir* — Span.-L2 *Buddy, that's my foot* WHICH *you're standing on*	*The people next door are rather* INDIGENT
Epenthesis	/sətreɪ/ for / streɪ/	—	—	—
Avoidance a) Topic avoidance 1. Change topic 2. No verbal response	(To avoid using certain sounds, like /l/ and /r/ in *pollution problems.*)	(Avoiding talking about what happened yesterday.)	(Avoiding talk of a hypothetical nature and conditional clauses.)	(Avoiding talk about one's work due to lack of technical vocabulary.)
b) Semantic avoidance	*It's hard to breathe* for *air pollution*	*I like to swim* in response to *What happened yesterday?*	Q: *¿Qué quieren los pájaros que haga la mamá?* R: *Quieren comer.* (Spanish-L2)	*Il regarde et il veut boire* to avoid the word for *cupboard* in *Il ouvre l'armoire*
c) Appeal to authority 1. Ask for form 2. Ask if correct 3. Look it up	Q: *f . . . ?* R: *fauteuil* (French-L2)	Q: *Je l'ai . . . ?* R: *prise.* (French-L2)	Q: *El quiere . . . ?* R: *que te vayas.* (Spanish-L2)	*How do you say "staple" in French?*
d) Paraphrase	*Les garçons et les filles* for *Les enfants* (Thus avoiding liaison in French-L2)	*Il nous faut partir* for *Il faut que nous partions* (To avoid subjunctive in French-L2)	*J'ai trois pommes* for *J'*EN *ai trois* (To avoid *en* in French-L2)	High coverage word: *tool* for *wrench* Low frequency word: *labour* for *work* Word coinage: *airball* Circumlocution: *a thing you dry your hands on*
e) Message abandonment	*Les oiseaux ga . . .* (*gazouillent dans les arbres* was intended in French-L2)	*El queria que yo . . .* (*fuera a la tienda* was intended in Spanish-L2)	*What you . . . ?*	*If only I had a . . .*
f) Language switch	*I want a* COUTEAU[2]	*Le livre de Paul's* (French-L2)	*Je ne pas* GO TO SCHOOL (French-L2)	*We get this* HOSTIE *from* LE PRÊTRE (English-L2)

A less error-analysis-based approach to the study of communication strategies is that of Faerch and Kasper (1983), which calls for "problem-orientedness" and "consciousness" as criteria for defining strategies: "Communication strategies are potentially conscious plans for solving what to an individual presents itself as a problem in reaching a particular communicative goal" (p. 36). Far from being a notion that can be ignored in a communicative analysis of a language test, the strategic component can offer us a way of analyzing test performance in terms of planning and problem solving. For example, it should be possible to devise multiple-choice problems of the TOEFL type that would allow the student to display an ability to (a) reach a desired communicative goal without problems, (b) implement an achievement strategy by choosing an alternative plan for reaching the desired goal, or (c) implement a reduction strategy by choosing an alternative goal. This notion offers exciting possibilities for test format development, and I hope to provide at this conference some examples of the type of strategy-oriented "receptive" test items I have in mind. In any case, it is my view that the area of interlanguage studies has insights to offer us in our analysis of testing in a communicative competence framework, and I would like to consider briefly some issues from an interlanguage point of view.

INTERLANGUAGE AND DISCOURSE DOMAINS

In the decade and a half since Corder's (1967) work on "The Significance of Learners' Errors" began the area of research that has come to be known as interlanguage (IL) studies, two aspects of the research seem to have relevance for our concern with communicative competence and testing: the recognition of variability in IL systems and an emphasis on the importance of interaction in IL studies. It has been fairly well documented that learners' IL systems vary with elicitation tasks and that the more careful the style, the closer to target-language norms the IL production is (Tarone, 1983). For example, Schmidt (1980) found in her study that learners from several native language backgrounds varied in employing second-verb ellipsis, such as "Mary is eating an apple and Sue X a pear," where X is the deleted verb "is eating." Schmidt's subjects never produced a sentence like the example in free conversation, but when asked to repeat it a few seconds after a model was presented, 11 percent could do so, and, when asked to combine two sentences with identical verbs, 25 percent employed second-verb ellipsis. Finally, 50 percent of the subjects correctly judged the form as grammatical in English. This simple example is presented as an illustration of the problem of characterizing a learner's competence when it would appear that "competence" varies with task. Below, I will suggest that this variation with elicitation task is a subset of variation with "context," a larger category, but that our communicative competence framework must include some mechanism for dealing with this kind of variation if it is to allow us to characterize what it means to know a language. If our tests can be thought of as "windows" on a learner's competence, then our competence paradigm must obviously be capable of interpreting for us what we see there.

Before discussing the question of context in IL studies, I would like to mention the other feature of recent work in IL, that of the emphasis on analyzing communication during interaction between speakers who do not share a first language. This emphasis is, of course, related to that discussed above on variation, since the greatest divergence from the target occurs typically in situations where the learner is interacting with a native speaker, focusing not on form but on communication. One can point to numerous recent studies of various aspects of interaction: studies of "foreigner talk," of classroom communication, of repair sequences, of tolerance for "nonnativeness," and so on. Certainly, too, pedagogical fashion has moved in recent years toward communication-oriented goals and methods, and in testing we are frequently being asked to provide more information about learners' abilities to use their IL competence in communicative situations--hence, part of the impetus for a conference such as this one.

I would like to suggest that our communicative competence paradigm needs to be evaluated in terms of its ability to deal with variation in production of target forms in different interactional situations, and to offer a discussion of an extension of the IL hypothesis proposed by Selinker and Douglas (1985) may provide a way to modify the communicative competence framework along the lines I have alluded to above.

As an example of the sort of phenomenon that Selinker and I have been attempting to investigate, I offer the following anecdotal evidence: we are acquainted with an applied linguist from Poland and have been listening to his Polish-English IL for several years now. We wish to suggest that he has created for himself (and he agrees with this) the following two "contexts" (which we will call "discourse domains"):

1. The first we will call "being an international professor who lectures in English."

2. The second we will describe as "telling stories about Poland, after drinking several vodkas."

In this regard, we should claim that in the first domain, that of being an international professor, our friend's use of the modal "should" in his English IL appears to be no different from ours, whereas in the second domain, that of telling stories about Poland, his use of the word "should" becomes quite different from ours: he repeats the modal more often in the discourse than we would. An example would be insertion of the word "should" in sentences with "if" clauses: "If Kristina should go to Warsaw, I should drink a toast." Our colleague does not produce sentences of this type in the professorial domain, but produces instead target-like "correct" utterances. The point is that, in order to describe out friend's English proficiency accurately, we cannot set up a cateogry "modal" and talk about "obligatory contexts" and the like. The modal "should" in the vodka/story-telling domain is linked to individualized narrative structure and temporal verbs that are not very English-like. Externally imposed categories seem to miss the point.

Recently, a number of studies of IL production within discourse domains have been conducted, examining particular ILs. One, Watanabe (1982), looked at the Japanese IL of an American missionary married to a Japanese woman, who had lived in Japan for two and one-half years and had been back in the U.S. for a year at the time of the study. The data were drawn from conversations between the researcher, a Japanese native speaker, and the subject in two areas that appeared important to the latter. One involved his work on behalf of Christianity, and the other involved the building of a storehouse, a most unusual endeavor for a foreigner in Japan and one of which the subject was clearly proud.

Watanabe notes several differences in the IL used in the two domains.

> The difference in the two discourse domains can...be observed in different grammatical usage: the use of the case markings is different in the two domains. In Japanese, all noun phrases are marked by case forms, which function to show grammatical relations (subject and object) as well as semantic relations (agent, patient, location, goal, etc.). They take the form of noun-case. In the domain of building, the subject tends not to use case markings, even after simple nouns, whereas in the domain of Christianity, he does use cases after simple nouns.(p. 5).

Watanabe's data show, for example, that in the Christianity domain, the subject provides the correct case markings in all thirty-eight obligatory environments after simple nouns (100 percent), whereas in the building domain, he does so only three times in twenty-three obligatory environments (13 percent). In a more complex context than simple nouns, the subject does the following: after relative clauses in the Christianity domain, he uses case marking in eleven of sixteen obligatory environments (69 percent), while, in the building domain, he does so only once in eleven obligatory environments (8 percent).

Another study that represents a recent attempt to gauge linguistic similarities and differences across IL domains is by Fakhri (1984). Fakhri looks at the problem of language use by a former American Peace Corps volunteer in Morocco who had not spoken Moroccan Arabic for four years at the time of the study. Fakhri audiotaped daily conversations with the subject in Moroccan Arabic during a four-week period to study changes and nonchanges in her English-Moroccan Arabic IL. He concludes that despite her long break with the target language (TL), "and evident linguistic deficiencies, the subject maintained to a certain degree a functional use of the TL." He notes that she was able to compensate for these linguistic deficiencies by using various strategies, which he calls "compensatory strategies." He suggests that discourse considerations provide an insight into the application of these strategies "within a specific mode of discourse." Fakhri looks at the constraints produced by her attempts at narrative versus procedural discourse. In the domain of relating her experiences in Morocco, she compensates for her deficient use of verb inflections by "using free morpheme pronouns in order to keep reference to participants straight." On the other hand, in the domain of describing procedures for cooking couscous in Morocco, she does not resort to this strategy but uses either pronoun inflections or the base form of the verb.

As a final example of variation related to discourse domains, and before returning to the question of the consequences of all this for the communicative competence framework, I would like to consider some data from Selinker and Douglas (1985). Our subject, whom we will call "Luis," was a Mexican graduate student in engineering at the University of Michigan. Conversations with the student in two domains--engineering and Mexican cooking--were videotaped and analyzed. The unit of analysis we attempted to employ was that of rhetorical units, such as describing an event, describing a process, defining a term, and so on. We referred to these units as "episodes" and attempted to look at pairs of similar episodes in the two domains to see whether the subject's IL varied with domain. The two episodes in this example will illustrate the subject's strategic ability to deal with missing vocabulary in the two domains. In the first engineering episode the interviewer had asked Luis why some construction equipment had broken down, referring to a text they were reading:

Luis: ...and then this is eff - eh - referring that the contractor maybe didn't adjust the equipment to the co - site conditions - maybe this you know the equipment can be affected by the the - what is that the - I lost the word - I mean - because no - for - you have one equipment here in for example one estate and you want to move that equipment to, for example, you are working Michigan and you want to move that equipment to Arizona or a higher estate - you have to adjust your equipment because the productivity of the equipment eh gets down - eh because of the different eh height of the (project) place

Interviewer 1: Oh, I see. the a the ah altitude

Luis: yeah, the altitude - that is the word =

Interviewer 1: That's the word

Luis: =I was looking for

In this episode, Luis is able to carry on in spite of the missing word. His strategy is to describe the process involved in moving equipment from one part of the country to another; he continues talking until he is able to access a synonym--height--for the forgotten word, which allows his interlocutor to suggest the correct technical term--altitude. In addition, as shown on a video tape of this episode, Luis does not resort to the use of gesture to get at the missing word--which would have been an easy thing to do in this case--but instead relies almost exclusively on verbal strategies.

In the matching episode, discussing Mexican cooking, Luis is talking with another interviewer and again forgets a word:

Luis: I don't know if you know what is - what machaca is.

Interviewer 2: Tell me - I think I've had it once before.

Luis: No - you you get some meat and you put that meat eh to the sun an' after that you - I don't know what is I - I learned that name because when I went to the sss - farmer jack I saw that - you make like a little thin - oh my god - then you - you - forget it
(laugh)

Interviewer 2: (laugh) Make it into strips?

Luis: OK. Lika a - you you have a steak no? You first

Interviewer 2: uh huh

Luis: In the sun - you have

Interviewer 2: Then it gets rotten and you throw it away.

Luis: Ummm - no, no, no, no, no, only one day or two days

Interviewer 2: Um hmmm.

Luis: After that with a stone you like escramble that like ah -

Interviewer 2: You grind it up?

Luis: Yes, that psss you you start to what is that word oh my god

Interviewer 2: Mash?

Luis: Exactly - you have to you start making mash that meat...

Although he is able to carry on in spite of the missing word, a more serious breakdown occurs in this nontechnical domain, where he produces the phrase "forget it." This contrasts with his performance in the technical domain, where no such breakdowns occur. Note that after the breakdown, he attempts the same strategy as in the technical domain--the strategy of describing a process, leading up to a synonym for the missing term, and then accepts the term from the interlocutor--but the outcome is not very successful. In the nontechnical episode, Luis struggles to remember the word, even mentioning that he learned it at a supermarket, "Farmer Jack," and attempts to break off the episode--"forget it." Only when his interlocutor suggests a wrong word does he begin to employ the communication strategy and describe the process of drying the meat and hitting it with a stone, waiting for the interviewer to offer another, correct term. The interviewer finally suggests "mash," which Luis accepts as the correct term. [NOTE: Selinker and Douglas (1985) suggest that in the nontechnical episode, the strategy failed in the end to achieve its communicative purpose. It is our belief that "mash" is not in fact the correct technical term Luis wanted to convey the meaning he sought--an action of crushing, grinding, and separating the dried meat--which is the process of making machaca. Probably no such word exists in English.] Finally, in this episode, the video tape shows Luis using a great number of gestures to assist him in overcoming his problem with the missing term--something he did not resort to in the technical domain.

My purpose in presenting these three sets of data concerning variation in IL use is to suggest that we need to adjust our ideas of the nature of communicative competence to take account of the sort of variation I have been describing. Tarone (1983) outlines three paradigms that have been put forward to represent the system underlying learner utterances: a Chomskian, homogeneous competence paradigm (Adjemian 1976), a capability continuum paradigm (Tarone 1979, 1982) and a dual competence paradigm (Krashen, 1981). Tarone argues that the capability continuum (a continuum of styles that produces variation in learner output according to the amount of attention being paid to form in any given situation) accounts for the variation in the data better than either of the other two paradigms. I do not suggest that the discourse domains concept, as Selinker and I have outlined it, should replace the existing formulations. However, I think the concept has potentially useful consequences for our understanding of communicative competence, and I wish to elaborate the notion briefly here and suggest how it might be integrated into the communicative competence framework.

There is a difficulty in moving from observable linguistic data to postulating the psycholinguistic mechanisms that underlie the performance, As Dennis Preston (1984) has pointed out, what is happening externally may not be what is happening inside, and weird psycholinguist theories can result. For example, to account for variation that produces, say, 75 percent of one form of a variable and 25 percent of another form during a single performance, one could postulate that an individual estimates how many instances of the variable will occur during the performance (using, presumably, information about past similar performances), monitors the number of occurrences of each form, and adds more of the first form when he or she senses that the 75 percent level is dropping off. One could also postulate racks of weighted coins, each rack appropriate to a certain set of contexts; each time the variable occurs, a weighted coin from the appropriate rack is tossed and, depending on whether the 75 percent or the 25 percent side comes up, one chooses the requisite form of the variable. One can, like Tarone (1983), postulate a style continuum, along which the learner shifts as attention to form increases or decreases. Or one can postulate, as Krashen (1981) seems to, and as I think Bickerton

(1973) does with his "dynamic paradigm," a set of largely independent systems that cause variation in output as a response to various environmental factors, the language user shifting rapidly between systems (that is, the rules aren't variable at all, but rather the rapid shifting gives the impression of variability).

I have grossly oversimplified a number of very complex hypotheses in order to make the same point Canale (1983) does: we need to begin to move in our thinking from the analytical framework provided so far by the communicative competence formulation toward a model of communicative competence that will specify how the competencies are acquired and how they interact in actual communication. Selinker and Douglas hypothesize that the important processes in IL learning--transfer, fossilization, backsliding, and the like--occur within discourse domains and have attempted to formulate a research methodology based on principles of ethnomethodology to study interaction between learners and native speakers in different domains. Our notion of discourse domain is that it is a personally and internally created area in the life of an interlanguage learner that has importance for the learner such that he or she must, or wishes to, interact in it. It is our claim that IL systems are acquired within these discourse domains, that learners' progress in acquisition must be studied with reference to domains, and, by extension, that teaching and testing might best take place with reference to them. I present below a vague notion of how discourse domains might be integrated into the communicative competence framework and help us out of a potential problem.

In his discussion of applications of the communicative competence framework for teaching and testing, Canale (1983) employs a metaphor of driving instruction to make the point that without training in communicative skills, in addition to training in communicative knowledge, our students will be ill prepared to handle actual communication. This point is well taken. However, to take up Canale's metaphor again--perhaps absurdly--for just a moment, imagine that the student drivers were not told that the knowledge and skills they were being taught were specifically for operating an automobile, but were for "general transportation." How likely is it that these students could then take their knowledge and skills of general transportation and apply them differentially to the operation of a car, a plane, or a steamship? Rather, each of these competencies is taught and learned independently. I would suggest that this point, applied to second-language learning, is not trivial, that communicative competence is developed differentially within different discourse domains, and that models of communicative competence will have to be developed domain by domain. Although this notion gets us no nearer to articulating the principles of interaction among the components of communicative competence, that is, toward developing a communicative model, as called for by Canale (1983), I believe that work will proceed much more fruitfully if discourse domains are taken to be the arenas of research. Principles of ethnomethodology, and research techniques developed within that framework (see Erickson, 1979; Frankel and Beckman, 1982; & Gumperz & Tannen, 1979) offer ways of studying communication within domains and also suggest ways of applying research findings to teaching and testing problems (for example, Douglas and Pettinari, 1983). In the sections that follow, I will consider how the TOEFL Listening Comprehension section and the Test of Spoken English might be changed to bring them closer to some aspects of communicative competence I have discussed here. In particular, I will examine the listening section from the point of view of authenticity, actual communication, context, and strategies. In this regard, I will refer to the development of a new listening comprehension component of the Michigan Battery at the University of Michigan. With the TSE, I will look at some of the texts produced by candidates responding to the section on describing a series of pictures from the point of view of discourse planning and development. In light of this discussion, I hope to make some suggestions for the TSE scoring procedure that might lead to more information about the candidate's communicative competence.

TOEFL LISTENING COMPREHENSION

Duran et al. (1985), in their analysis of the TOEFL in a communicative competence framework, criticized the Listening Comprehension section in four general areas: (1) the items in all three parts were not representative of authentic language use, exhibiting as they do crisp enunciation, minimal ellipsis, and no unexpected pauses or false starts; (2) the items, especially in Part B, did not exhibit such characteristics of actual communication as negotiation of shared meaning, coparticipants working together to shape and direct the flow and structure of the text; (3) the items failed to provide much sociolinguistic information; and (4) the items failed to exhibit such strategic qualities as hesitation, repair, unfinished statements, paraphrasing, rephrasing in midsentence, and side comments. These are worthwhile criticisms, and I have little to add to them, except that it is obvious that inclusion of a strategic competence component in the evaluation instrument would have added to the value of the analysis by Duran and his colleagues. I would like to offer some suggestions for approaches to test construction that may provide a basis for answering these criticisms (though not all at once), based primarily on the work going on at the University of Michigan English Language Institute in producing a new listening comprehension section of the Michigan Battery.

With the express aim of making the Michigan listening section more communicative, ELI testing staff have been experimenting with a new format roughly analogous to the TOEFL Listening Comprehension section, Part C, involving minitalks and extended conversations, but with some, I think, crucial differences. Briefly, there are two subtests in this part of the listening test: a two- to three-minute minilecture and a three- to four-minute dialogue. Each language sample is followed by ten to fifteen multiple-choice problems to make a total of twenty-five for this part of the test. Features that add to the communicative nature of this part of the test are (1) the speech samples display authentic qualities, including hesitation, repair, paraphrasing, and the like as well as natural enunciation, ellipsis, false starts, and pauses; (2) the dialogue exhibits some features of interaction such as negotiation of the structure and flow of the text; and (3) both texts involve information transfer between aural and visual modes, in the form of diagrams, charts, tables, or maps that are referred to in the texts. The central task subjects must accomplish is the interpretation of the visual information, and the questions refer mainly to this. Subjects are invited to make whatever notes they wish, directly on the test paper, to assist them in answering the questions, which are presented orally.

Many of the authentic features of the texts are a result of the production methodology. In the case of the minilecture, the presenter is asked to deliver the talk either extemporaneously or from skeletal notes, not by reading it from a written version. Often the visual material--the chart or graph--is the starting point of the lecture. This initial text is taped and the recording transcribed. The transcript is analyzed for structure and content, essentially to ensure that there is a logical development and that references to the visual material are clear and unambiguous. The text is discussed with the presenter and changes are agreed upon. An outline of the text is prepared and acts as a guide to the presenter in rerecording the talk. The testing staff may go through this process of recording, analysis, and rerecording several times before an acceptable text is developed. The aim is to achieve a text that, while containing a well-formed underlying structure, nevertheless exhibits features of authentic language production.

The dialogue production process is somewhat similar to that of the minilecture, but with a few interesting differences. The two speakers decide in general terms what they will talk about--for example, the location of a picnic spot or the rearrangement of the furniture in a room--and then record the conversation quite spontaneously. It is in this first recording that the negotiation of structure and flow takes place. In fact, even in the later versions one hears the voices overlapping--the speakers echoing each other, correcting each other, and generally displaying features of actual communication.

Unlike the minilecture, the dialogue production involves the creation of the visual information during the transcription and analysis stage. For example, if the speakers are discussing the rearrangment of furniture, a diagram of the room with the furniture in it will be prepared. Clearly, the function of the visual material is quite different in the two subtests: in the minilecture, the chart or diagram serves as a reference point for both the lecturer and the listeners, while, in the dialogue, the visual material is never referred to by the speakers, since they already share a mental image of the topic of conversation--the room or the picnic spot. The function of the diagram or map in this case is to provide this "prior text" for the listeners who do not share it.

It seems to me that this development procedure has something to offer researchers interested in producing tests of communicative competence, allowing as it does for the controlled construction of authentic texts that can involve the subjects in problem-solving communicative activities. More information of a sociolinguistic nature should probably be provided, particularly in the dialogues, so that subjects could be tested on their sensitivity to appropriate language use as well as on their abilities to cope with factual information and to make transitions from aural to visual data. A great deal of analysis remains to be done on test formats of this type to determine their construct validity, particularly in a communicative competence framework, since they are so complex in discourse and strategic terms. It is important to experiment with texts of this type, which exhibit an underlying controlled regularity, while at the same time displaying a surface "roughness," since in this way we can give subjects an opportunity to display their ability to "make sense" out of authentic language events. I would like to turn now to a discussion of the Test of Spoken English and look at one section of it from the point of view of strategic competence, in order to make some suggestions for scoring procedures in the program.

THE TEST OF SPOKEN ENGLISH

Section 4 of the TSE requires subjects to tell a story based on a set of six pictures. Although this task is fairly controlled, in that all information is provided in the pictures, which are of a very simple nature, there is considerable opportunity for subjects to exhibit various planning and execution strategies for accomplishing the task. An analysis of responses to Section 4 could perhaps provide us with information that would be useful not only in evaluating subjects' performances but also in developing further a concept of a communicative competence model. As an example of this type of analysis, I have looked (not very rigorously, in this first attempt) at texts produced by two Korean graduate students who took the Speaking Proficiency English Assessment Kit (SPEAK) (Test of Spoken English, 1982) version of the TSE at Wayne State University in August 1984. One (Text 1) scored a 240 on the test, while the other (Text 2) scored 110. Following is the unsegmented version of each of the texts:

TEXT 1

One day last month I rode bicycle it was locked so I unchained it and I rode the bicycle wh... during..I..while..I ride bi...riding bicycle I saw the girl who was riding bicycle wh... I was keep looking at that girl because she was pretty i... somehow I miss the road I hit the tree I broke the leg so I had t' go to the hospital while I was there...she...visited me with flowers a...that helps me...cure..my leg..fast...after that we begin...ge...to... be a friend and we rode bicycle every time we have... any chance

TEXT 2

One day last month I'd like to ride a bike and then ah so I open my key which which is was which lock bi....locked bike and then...I was riding a bike but my girlfriend passed my my way so I looked at ah my girlfriend and then at that time I crashed the tree so I broke I broke my leg and then so my f...my girlfriend come to my my ah ho... hospital and then and so an that time I and my my girlfriend an take a bike

The first text contains 100 words (not counting incomplete words), while the second contains 89 words. Segmenting the texts relative to the picture sequence, we get the following:

TEXT 1

Picture 1: One day last month I rode bicycle it was locked so I unchained it

Picture 2: and I rode the bicycle

Picture 3: Wh...during...I...while...I ride bi...riding bicycle I saw the girl who was riding bicycle uh I was keep looking at that girl because she was pretty

Picture 4: I... somehow I miss the road I hit the tree

Picture 5: I broke
the leg so I had t' go to the hospital while I was there... she...visited me with flowers a...that helps me...cure..my leg..fast...

Picture 6: After that we begin...ge...to... be a friend and we rode bicycle every time we have...any chance

TEXT 2

Picture 1: One day last month I'd like to ride a bike and then ah so I open my key which which is was which lock bi....locked bike

Picture 2: and then...I was riding a bike

Picture 3: but my girlfriend passed my my way so I looked at ah my girlfriend

Picture 4: and then at that time I crashed I crashed the tree

Picture 5: so I broke I broke my leg and then so my f...my girlfriend come to my my ah ho... hospital

Picture 6: and then and so an that time I and my my girlfriend an take a bike

The two subjects vary in the number of words they produce per segment:

Segment	TEXT 1	TEXT 2
1	14	25
2	5	7
3	26	13
4	9	11
5	27	18
6	19	15

The first subject clearly finds more to say about pictures three and five of the SPEAK test than does the second subject, noting that the girl was pretty and that she brought him flowers in the hospital, which helped cure his leg fast. Another way of segmenting the texts is by reference to falling intonation and pauses. Doing so, we get the following:

TEXT 1

1. one day last month
2. I rode bicycle
3. it was locked so I unchained it
4. and I rode
5. the bicycle
6. wh... during I while I ride bi riding bicycle
7. I saw the girl
8. who was riding bicycle
9. uh I was keep looking at that girl
10. because she was pretty
11. I... somehow I miss the road
12. I hit the tree
13. I broke the leg
14. so I had t' go to the hospital
15. while I was there
16. she...visited me with flowers
17. a...that helps me...cure..my leg..fast...
18. after that we begin...
19. ge...to...
20. be a friend
21. and we rode bicycle
22. every time we have...any chance

TEXT 2

1. One day last month
2. I'd like to
3. ride a bike and then ah so
4. I open my key which which is was which lock bi....locked bike
5. and then
6. I was riding a bike but

7. my girlfriend
8. passed my my way
9. so I looked at ah
10. my girlfriend
11. and then
12. at that time
13. I crashed I crashed the tree
14. so I broke
15. I broke my leg and then so my f...
16. my girlfriend
17. come to my
18. my
19. ah ho... hospital and then
20. and so
21. an that time
22. I and my my girlfriend an' take a bike

This type of segmentation begins to give us a bit of insight into the planning and execution process. Note that the units sometimes overlap the picture boundaries, suggesting that the subjects are structuring the text according to their own plans, not bound strictly by the picture sequence. More interestingly, sometimes the subjects produce strings that are very native-like in their structure:

TEXT 1

it was locked so I unchained it
I saw the girl
because she was pretty
I hit the tree
so I had t' go to the hospital
while I was there

TEXT 2

I was riding a bike
so I looked at
at that time
I broke my leg

We can hypothesize, after Lesser and Erman (1977) and others, that strings of this type represent "islands of reliability" or "anchors" in the flow of speech that facilitate the planning and execution of the discourse. The supposition is that the speaker plans for these "islands" and inserts them into the discourse partly to give himself a base from which to search for, test, and structure linguistic units to make up more of the discourse. Viewed in this light, the notion of communicative strategies takes on a much more central role in the communicative competence framework and may offer, ultimately, a clue to the processing model for which we continue to search.

I have yet to offer suggestions on improving the communicative nature of the TSE evaluation system. The scoring criteria of grammar and pronunciation are obviously related to linguistic competence, and that of fluency, probably, though less clearly related to strategic competence, while that of overall comprehensibility is hard to classify, but seems most related to linguistic competence. The test tasks are hard, too, to classify from a communicative point of view: there is an oral reading task, a sentence completion task, story-telling based on a series of pictures, answers to questions based on a single picture, open-ended answers to questions of "international import," and, finally, an oral restatement of a written class schedule. While all of these tasks can be seen to have certain elements of communicative competence, only the last bears any firm relationship to what we might think of as actual communication, specifying as it does audience, setting, and purpose. However, what communicative aspects there may be in the tasks are largely lost in the scoring procedure, which requires scorers to ignore content and focus on primarily linguistic features of the responses.

Now, it would certainly be an easy matter to be an armchair test critic and say "change this" and "change that" in the TSE program. Having scored a fair number of SPEAK tapes, I can testify that the scoring procedure is quite workable and does seem to produce an evaluation that bears some relationship to what one intuitively feels to be a candidate's strengths and weaknesses in oral production. One would not want to suggest a number of exotic "improvements" that would, in fact, be unworkable in any sort of practical situation. Further, in my experience, the candidates themselves seem to feel that the test measures something corresponding roughly to their abilities to use English orally. Still, much of the candidates' performances is lost in the evaluation process, which is necessarily an abstraction. In order to remedy this I return to the central theoretical discussion in this paper, involving discourse domains, and suggest that by moving in the direction of domain-specific tasks, the TSE (and for that matter, the TOEFL) could be made to conform more to a framework of communicative competencies.

In the discussion of discourse domains, I suggested that the notion of discourse domains may have a central place in theories of second language acquisition and communicative competence, that second languages may in fact be learned within discourse domains, and that the important processes of interlanguage take place within them. I went on to suggest that the notion could have important consequences for our thinking on test development and score interpretation. Its importance is this: the specification of the discourse domain with respect to each task in a test will provide the scorers with a handle, a frame of reference within which to interpret the performance of the candidate. As Oller (1984) argues so convincingly, "...to the extent that the facts are fixed in any communicative activity, and to the extent that the criteria [for tests of communicative competence] are met, the task will be scalable in a meaningful way" (p. 42). It may be that domain specification will have such an effect at least partly because it will enable the test candidates to more efficiently plan for the "islands of reliability" referred to in the TSE discussion above, thereby providing a more coherent framework for the evaluators to judge their linguistic performance. As a first step in moving toward such specification, I would suggest that more information be provided to candidates about the audience, setting, purpose, and the like of the test tasks and that an additional scoring criteria of something like "appropriateness of response" be added, at least in some sections. This would go some way toward providing a communicative diagnostic in evaluating test performance, something lacking in the present formulation.

CONCLUSION

In this paper I have done three things: (1) outlined a formulation of discourse domains that may be integrated into the communicative competence framework as a first step toward producing a communicative model; (2) made procedural suggestions for listening test production in the TOEFL program; and (3) looked at one section of the TSE from a discourse perspective and suggested changes in evaluation procedures that brought me back to the discourse domains concept. Though I have made very few concrete suggestions for improving the communicative basis of the TOEFL and the TSE, I would hope that one or another of the ideas presented here might suggest to colleagues approaches to planning and carrying out research and development activities, taking us closer to the goal of a model of communicative competence, and also toward the goal of better evaluating learners' communicative abilities.

REFERENCES

Adjemian, C. (1976). On the nature of interlanguage systems. *Language Learning*, *26*, 297-320.

Bickerton, D. (1973). Quantitative versus dynamic paradigm: The case of Montreal 'que'. In C. Bailey & R. Shuy (Eds.), *New ways of analyzing English* (pp. 23-43). Washington: Georgetown University Press.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-27). New York: Longman.

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

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics*, *5*, 161-170.

Douglas, D., & Pettinari, C. (1983). *Psychiatric interview training modules: A program in language and culture for foreign medical graduates*. Paper presented at TESOL 1983, Toronto.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. E. (1985). *The TOEFL from a communicative viewpoint on language proficiency: A working paper* (TOEFL Research Report 17). Princeton, NJ: Educational Testing Service.

Erickson, F. (1979). Talking down: Some cultural sources of miscommunication in interracial interviews. In A. Wolfgang, (Ed.), *Nonverbal behavior* (pp.99-126). New York: Academic Press.

Faerch, C. & Kasper, G. (1983). Plans and strategies in foreign language communication. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 20-60). New York: Longman.

Fakhri, A. (1984). The use of communicative strategies in narrative discourse: A case study of a learner of Moroccan Arabic as a second language. *Language Learning*, *34*(3), 15-38.

Frankel, R., & Beckman, H. (1982). IMPACT: An interaction-based method for preserving and analyzing clinical transactions. In L. Pettigrew (Ed.), *Explorations in provider and patient interactions* (pp. 71-85). Nashville: Humana Press.

Gumperz, J., & Tannen, D. (1979). Individual and social differences in language use. In C. Fillmore, D. Kempler, & W. Wang (Eds.), *Individual differences in language ability and language behavior* (pp.305-325). New York: Academic Press.

Krashen, S. (1978). Second language acquisition. In W. O. Dingwall (Ed.), A survey of linguistic science (pp. 317-338). Connecticut: Greylock Press.

Krashen, S. (1981). Second language acquisition and learning. Oxford: Pergamon Press.

Lesser, V., & Erman, L. (1977). A retrospective view of the Hearsay II architecture. Proceedings of the 5th International Joint Conference of Artificial Intelligence, Volume 2, (pp. 790-800). Cambridge, MA: Massachusetts Institute of Technology.

Oller, J. W. (1984, October). Communication theory and testing: What and how. Paper presented at Second TOEFL Invitational Conference, Princeton, NJ.

Preston, D. (1984). Personal communication.

Schmidt, M. (1980). Coordinate structures and language universals in interlanguage. Language Learning, 30(2), 397-416.

Selinker, L., & Douglas, D. (1985). Wrestling with 'context' in interlanguage studies. Applied Linguistics, 6(2), 190-204.

Tarone, E. (1979). Interlanguage as chameleon. Language Learning, 29(2), 181-191.

Tarone, E. (1982). Systematicity and attention in interlanguage. Language Learning, 32, 1.

Tarone, E. (1983). On the variability of interlanguage systems. Applied Linguistics, 4(2), 142-163.

Tarone, E., Cohen, A., & Dumas, G. (1983). A closer look at some interlanguage terminology: A framework for communication strategies. In C. Faerch & G. Kasper, (Eds.), Strategies in interlanguage communication (pp. 4-14). New York: Longman. Originally published in Working papers on bilingualism, 1976, 9, 76-90.

Terrell, T. D. (1980). The natural approach to language in language teaching: An update. Irvine, CA: University of California. Mimeo.

Watanabe, N. (1982). A study of English-Japanese interlanguage. Unpublished manuscript, University of Oregon.

Test of Spoken English. (1982). Speaking Proficiency English Assessment Kit. Princeton, NJ: Educational Testing Service.

A RESPONSE TO JOHN OLLER'S "COMMUNICATION THEORY: WHAT AND HOW" AND TO DAN DOUGLAS'S "COMMUNICATIVE COMPETENCE AND TESTS OF ORAL SKILLS"[1]

Frances B. Hinofotis

I found both John Oller's paper and Dan Douglas's paper to be extremely provocative because each suggests practical test development issues that need discussion and empirical investigation. I have chosen in responding to these papers to look at three areas as general language testing concerns and as they relate specifically to the TOEFL. I will come to them in a moment.

I want to preface my remarks by saying that underlying my discussion is the desire that we all hope to see language assessment move in the direction of more effective procedures for tapping communicative ability across the skill areas. Any advancement in that direction will depend in large part on evolving theoretical perspectives based on careful research that will help us better understand what it means to be communicatively competent. John Oller, in his proposed theoretical framework, suggests that communicative competence and general intelligence can be viewed as essentially the same thing. This psychological perspective is intriguing, and I look forward to seeing where it, along with Candlin's sociological model, will lead the field. Dan Douglas, taking the Canale and Swain (1980) framework as a point of departure, urges us to look carefully at our communicative competence paradigm in terms of the variability in interlanguage systems. This variability, he demonstrates, is often reflected in different interactional settings or discourse domains.

As a starting point for my own observations on assessing communicative competence, I want to consider the notion of context. I will then mention authenticity as it relates to test content. Finally, I will look at "interestingness" or interest value of a text within the framework of a test. In all three areas, I will be drawing on points that Oller and Douglas made, and I will be raising questions more than I will be providing answers. I hope that some of the questions I raise will work themselves onto the long-term research agenda of the TOEFL program.

One thing that may be immediately clear is the interrelationship among the three topics I will be discussing. Test development decisions in one area will have consequences in the other two. It is often difficult, for example, to separate authenticity and interest level when trying to determine whether to use a passage or a piece of oral discourse in a test because part of what may make one or the other interesting is its authenticity, its relevance to the test taker.

Context

Context in testing, it seems to me, can be considered in at least two ways: (1) the context in which a test is to be used, where concerns such as purpose and target audience in a general sense are initial considerations, and (2) the context for individual items or a series of items where the context will take different forms depending largely on the skill area(s) to be tested within the broader context established first. Underlying any decision about what type of test will be developed or selected in a given situation is the answer to the basic question, "What kind of information do we hope to obtain about a person's proficiency in some area--in our case, language proficiency, and, even more specifically, communicative ability with the language?"

With the TOEFL we are looking at language proficiency in a specific context, an academic context, and for test development purposes we need to keep that in mind as we consider the emerging broader framework of communicative competence as a whole. English for academic purposes (EAP) is by definition English for a very specific purpose: survival in an English-speaking college or university environment. Conceptually, there is nothing new here. But there is a problem or at least a challenge facing the TOEFL program, hence the reason for this conference. That challenge involves specifying as fully as possible what it means to be communicatively competent enough in English so that language will not in itself prevent a person from succeeding at an American college or university.

Given that the TOEFL is a general EAP test in that it cuts across disciplines and levels of academic study, how accurately do the tasks on the TOEFL reflect what students have to do on a day-to-day basis with the language? There are certain kinds of tasks, such as following written and oral directions, asking questions for clarification, answering questions orally and in writing, understanding lectures, and understanding reading assignments, that are classroom and study oriented. Others, such as having to select classes and register, having to meet with advisers, having to obtain information in the library, are of a campus survival nature. Still other areas come to mind, such as arranging for housing, eating, and transportation, that are more of a general survival nature but are clearly within the realm of day-to-day experiences for a foreign student. Somehow these all help characterize, in a general sense at least, part of the skills necessary to survive in an academic environment.

The general survival skill areas are probably not areas of critical concern for the TOEFL, certainly not to the extent the clearly academic areas are, though it could be argued that tapping those areas would provide a more complete picture of the potential student's communicative ability. But, regardless of ultimate scope of the total context, tasks must be identified and fully specified so that test developers can move in the direction of producing authentic items.

The TOEFL program recognizes the need for fuller specificity, and some steps have already been taken in that direction. The work of Bridgeman and Carlson (1983), which surveyed the academic writing tasks of graduate and undergraduate foreign students, provides the groundwork for full specification of academic writing tasks. Follow-up work in this area, which includes experimentation with formats and item types, could lead to a Test of Written English (TWE) to complement the TOEFL in writing as the Test of Spoken English (TSE) does in speaking. Oller (this volume) indicates that development of a TWE might be advised. I concur.

Work is currently underway by Don Powers on identifying listening comprehension skills needed by university students. This research is another positive step toward a fuller specification of what it means to be communicatively competent in English in an academic context. This attention to listening comprehension needs will help to fill the gap in communicative competence research regarding receptive skills, something Douglas mentioned in his paper. Listening comprehension tests are notorious for being generally less reliable than other types of language tests. Research that could help us better understand the complexities of listening processes will surely lead to the development of more effective listening tests. One direction for research that I think has potential in this regard, just as Douglas does, is that of looking at strategies nonnative speakers employ when their language per se, their linguistic competence, fails them. To what extent is compensation possible in various situations to help them achieve their goals? How can we identify and then test for repertoires of strategies?

Kathi Bailey (1983), in her dissertation research on foreign teaching assistants (TAs), looked in part at how some TAs were more effective than others at compensating for both productive and receptive language weaknesses. Their abilities to compensate, given relatively equal language proficiency ratings, had a strong effect, as you might expect, on their effectiveness as TAs. This same ability to compensate for language weaknesses is operating with nonnative students in other situations as they go about their academic tasks. Douglas cited Fakhri's 1984 paper, in which the term "compensatory strategies" is used. That paper, in addition to Douglas's paper for this conference, could serve nicely as points of departure for a researcher interested in this area.

Before I leave the topic of context, I want to look briefly at the notion of discourse domain that Douglas introduced in his paper. I think the notion of discourse domain--"a personally and internally created area in the life of an interlanguage learner that has importance for the learner such that he/she must or wishes to, interact in it" (p. 12)--is compatible with and, in fact, enhances the notion of specificity of tasks. However, one of the features of the discourse domain approach that is most appealing to me is also one that I perceive as being potentially problematic as we attempt to apply the concept to testing, and that is the personal nature of discourse domains.

As professionals, I think we have long realized the importance of having the learner's needs and interests reflected in teaching situations and testing situations. In Interagency Language Roundtable oral interview testing, for example, the testers to a large extent tailor each interview to the background and interests of the person being tested. It is more difficult to accomplish this in large-scale testing because of the practical restrictions we operate under. The TOEFL program cannot usually focus on individuals per se in test development, but rather must think in terms of group similarities (though to some degree the TSE allows for individuality in answers, and certainly a TWE would too). To the extent overlap in interests and needs can be identified, or--putting it another way--to the extent common discourse domains can be described, TOEFL test developers can incorporate the information in test specifications. I wonder to what extent an academic environment can be considered a broad discourse domain. I wonder, too, to what extent a testing situation is itself a discourse domain.

Authenticity

Next, I want to consider briefly the concept of authenticity in language testing. I will maintain the distinction Widdowson (1978) makes between genuine and authentic, with "genuine" referring to the use of real-world materials, and "authentic" referring to the learner being required to deal with those materials in "a way which corresponds to his normal communicative activities" (p. 80).

I would argue that it is easy enough in most cases for us to identify or produce genuine language samples for inclusion on a test. In an academic environment, we can select journal or textbook materials that are appropriate to a given group for assessing reading skills. We can tape-record lectures or discussions or generate them ourselves. Following some of Douglas's suggestions based on work at Michigan, these samples could be very natural from an auditory point of view. But given our concern for assessing communicative competence, the real question is, how authentic are the tasks we ask students to perform with the materials?

One step toward assuring that tasks are to some degree authentic is by doing what Oller (1984) calls "fixing the facts." It seems to me that "anchoring the factual domain" (p. 42), that is, using genuine material in an authentic way, is essential for test takers to feel that what they are doing has any resemblance to real-life activities that are important to them. How do we as test developers do this? Fixing the facts so that the content seems natural isn't always easy to do. Often, however, I think we can move much closer to fixing the facts by framing test items so that they at least reflect an authentic activity. For example, in assessing reading comprehension in an academic environment, instead of providing isolated reading passages, however geniune, with a series of questions following each one, the test taker could be told, in the directions or in some sort of lead-in, that the passages are being read for a specific academic purpose with the questions asked being somehow geared to that purpose.

I think a question that is often in the minds of test takers is, "Why am I being asked to do this?" Careful attention to the specificity of tasks on a test vis-a-vis real-world contexts (or discourse domains) and then careful fixing of the facts to frame the tasks should help make it unnecessary for the test taker to wonder why he or she is being asked to do something on a test.

Interest Level

Finally, I would like to consider the issue of interest level of a written passage or piece of oral discourse. It is clear from Oller's paper that he shares this concern. The issue is one that everyone involved in test development thinks about and wrestles with. A major question is how can we assure the selection of texts that are not only genuine but interesting as well--but, interesting by whose standards? What sort of guidelines can we provide for item writers to help them along these lines with their tasks? The major difficulty, of course, is that interest, generally speaking, is relative. Case in point: in discussing Listening Comprehension, Section 1, Part C, of the TOEFL, Oller states that the Ellis Island example minilecture is long and deadly dull, and then he asks, "Who cares about Ellis Island?" (p. 53). Immigrants and their descendants do. I care about Ellis Island. All four of my grandparents came through Ellis Island. People interested in social and cultural history probably care about Ellis Island. The lecture may be long and dull, not because the topic is Ellis Island, but rather because the information is presented poorly and perhaps not the most interesting facts were chosen. Oller, of course, exaggerated to focus our attention on the issue of interest level of a text within a test. My example is also something of an exaggeration, but I do feel that interest level affects test performance in some undefined way.

There is, however, an opposite perspective that test developers should consider.[2] The question here is, "Is there a danger of being overly concerned about the interest level of a test?" Consider the following statements:

1. We are not developing tests to entertain test takers.

2. Test takers expect a certain amount of dullness.

3. Some of the real-life language we have to deal with is uninteresting.

4. A testing situation is a special, real-life social setting (perhaps a discourse domain, as I suggested above), often loaded with pressure and anxiety, and this fact may override what is interesting.

5. If a passage is too interesting it may distract the test taker from the immediate task.

All of these points notwithstanding, for they are pragmatic and timely, I feel that "interestingness" is an issue we must continue to consider. I think there is enough uncertainty about the role of interest level in testing to warrant careful research that examines the issue systematically.

In closing, I want to return to a point I made earlier, and that is the importance of recognizing the interrelationship of context, authenticity, and interest level. It may be that as we move toward more fully specifying the tasks nonnative speakers must carry out in English, test content across the skill areas will become less problematic. With the TOEFL, the clearer it becomes what it means to be communicatively competent in English in an academic environment, the more effectively TOEFL test developers will be able to tap specific communicative skills. Both Oller and Douglas in their papers for this conference have offered direction in this regard. Oller urges us to "fix the facts" as we develop test items, to make clear to the test taker the context and relevance of the task. Douglas suggests that, by identifying discourse domains, we will be able to more clearly focus on the test-taker's needs with English. These suggestions, as well as others that have emerged during the past two days, provide impetus for modifying and expanding the TOEFL to better serve the test taker and the score user.

Notes

1. I wish to thank Colin Churchill, Thom Hudson, Brian Lynch, and Susan Stern for their comments on an earlier draft of this paper. I would also like to thank them for many stimulating discussions that contributed to my thinking on the issues discussed here.

2. I am indebted to Thom Hudson and Brian Lynch, doctoral students in the Applied Linguistics Program at UCLA and part-time colleagues of mine at National Education International, for offering this perspective.

References

Bailey, K. M. (1983). Teaching in a second language: The communicative competence of non-native speaking teaching assistants. Unpublished Ph.D. dissertation, UCLA, Applied Linguistics Program.

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

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

Candlin, C. N. (1986). Explaining communicative competence. In C. W. Stansfield (Ed.), Toward communicative competence testing: Proceedings of the Second TOEFL Invitational Conference. Princeton, NJ: Educational Testing Service.

Douglas, D. (1986). Communicative competence and tests of oral skills. Toward communicative competence testing: Proceedings of the Second TOEFL Invitational Conference. Princeton, NJ: Educational Testing Service.

Fakhri, A. (1984). The use of communicative strategies in narrative discourse: A case study of a learner of Moroccan Arabic as a second language. Language Learning, 34(3), 15-37.

Oller, J. W., Jr. (1986). Communicative theory and testing: What and how. Toward communicative competence testing: Proceedings of the Second TOEFL Invitational Conference. Princeton, NJ: Educational Testing Service.

Widdowson, H. G. (1978). Teaching English as communication. London: Oxford University Press.

DISCUSSION OF THE PAPERS BY OLLER AND DOUGLAS AND THE RESPONSE BY HINOFOTIS

Richard Tucker. Let us begin this discussion by allowing John and Dan to respond to Frances's reaction to their papers.

John Oller. I find myself in substantial agreement with what both Dan and Frances have said. I think the sample tape containing unrehearsed conversational speech that Dan brought with him, in spite of its inadequacies, shows that one needs to pay attention to the facts, that is, what one needs to know about a context, and take that into account in the construction of test items.

Another comment I would like to make concerns authenticity. Some years ago in Canada I was involved with a group from the Quebec Ministry of Education that produced language tests for secondary schools. The group produced a huge number of tests, some thirty-seven per year, and they had an extensive amount of item analysis data. It consistently showed that certain types of items were obviously better than the others. One of the features of the better items was what we have been characterizing as authenticity and interestingness. One such item type was a cartoon strip where there is a meaningful action going on, and then there is a conclusion with a punch. Such cartoons worked much better than cartoons without the punch. One study that examined that sort of thing with reference to standardized tests was the study that Engelskirchen, Cottrell, and Oller (1981) did at the University of New Mexico. We looked at the naturalness and fit of items that appeared on the Ilyin Oral Interview (Ilyin, 1976). That kind of design could be used in respect to a study of the TOEFL. You can rate the items in terms of their naturalness, and then compare the performance of the items, in terms of traditional item statistics, with the ratings. I think in this way you could work out a good practical solution to the problems of authenticity and interestingness.

Dan Douglas. Frances raised the issue that the personal aspects of discourse domains could be a problem, and I certainly agree with her. In his paper on form procedures, Larry Selinker (1979) arrived at certain conclusions about the speech of an individual he studies. Huckin and Olsen (1984) tried to replicate the same procedure with the same individual and arrived at different conclusions. Frances asked three questions relating to the size of the discourse domain. She wondered if it is the entire academic environment or if it was just limited to a single subject, such as physics. This is something we need to talk about. She wondered if tests are a discourse domain. I suspect that they are, but the consequences of that are pretty astounding. Her discussion of interestingness reminded me of some work that Tim Johns (1981) did a few years ago, in which he found that the comprehension of lectures per se by foreign students was not a problem. However, the comprehension of certain lecturers was. This shows us that people do things differently and the students' task is to respond to those differences. I think we are only beginning to realize the importance of the personal aspects of discourse domains.

Charles Stansfield. Both John and Frances say that the interest level of the examinee will be maintained if the text is well written. That poses a perplexing problem for us. For years the TOEFL Committee of Examiners has been debating whether or not it should "clean up" the language it finds in source materials, which are usually college textbooks written by professors of particular subjects. Frequently, this language is ineloquent. If we polish it, then it loses its authenticity, but if we don't, the propositions that the authors wish to convey remain unclear. This is a real dilemma for testers.

Lyle Bachman. I agree with Frances that context, authenticity, and interest level are interrelated. Returning to her question as to whether tests represent a discourse domain, I firmly believe that they do, and this should affect how we look at the issue of authenticity. I think test taking is similar to reading literature. Reading literature is a discourse domain. When we read literature, we need to willingly suspend our disbelief in order to enjoy it as a piece of aesthetic work. To look for complete authenticity in tests would be like searching for the pot of gold at the end of the rainbow. The most we can hope for is to get our students to willingly suspend their disbelief and play the game of taking the test as if it were real language. This is a realistic view of test taking that reinforces the point that it represents a specific domain of language use.

Regarding the authenticity of information sources, I think that interest is certainly a factor in authenticity, but I don't think it is the most important factor. A lot of things affect authenticity: one is artificiality---the extent to which the discourse satisfies Grice's (1975) maxims. I think level of difficulty is a crucial factor in level of authenticity. I don't care how interesting the text is; if it is too difficult for a student, he or she can not interact with it authentically. So I personally think we are placing too much emphasis on interest. I think we ought to be talking about authenticity and all the different things that can affect authenticity.

Russell Campbell. I don't think we need worry about the interestingness of the text. The fact that students are taking a test will be motivation enough for them to give their full attention to the text.

Chris Candlin. I think we may be getting into a bind here. If we associate interestingness with authenticity, then we cannot say what is authentic for a particular testee. The problem is that people "authenticate" the same situation in different ways. All social-psychological research shows that people have different perspectives on the same events. To assume that a text was authentic for a particular testee would mean that we would have to be able to deliver a text and a task to an individual in keeping with his or her particular perception of the events in question.

Let me relate this to the ESP issue, because this is a similar world that has been fully explored. Most of the work was done in Europe during the late 1970s. Research conducted on the subject was very specific as to the academic domain that was being described. The problem was that you could never get specific enough. Engineering wasn't specific enough, mechanical engineering wasn't specific enough, and branches of menchanical engineering weren't specific enough. In fact it was the paradox of the Munby (1978) model that if you really carried it out correctly you ended up with specificity at the level of the individual.

This morning Frances very interestingly talked about authenticity of task. What we need to do is to move up a level to some general academic tasks that all students in a wide variety of areas perform. Let's consider what these might be. They would certainly include textual manipulation of some grammatical type. They would clearly include the ability to handle propositional coherence within an entire text. They might also include various cognitive operations, such as the capacity to select relevant information and order it and reformulate it on one's own terms. If we can determine what these general academic tasks are, anything we put on the test in term of some academic lexicalization is a kind of academic wrapping designed to give the testee a flavor of the academic world.

Now it may well be that these central academic tasks relate to text, to coherence, to collation, and things of this kind. Tasks like finding one's way around the library have no place on this list. The only way I would use such a task is as a metaphor for collecting information and organizing it. So, if we go down this authenticity road, we will end up with a list of great magnitude. We should really move up to the level of basic textual competence, propositional coherence in connected text, various cognitive operations such as collating information, and relating this to some general academic form of lexis. But please let's not go down this ESP road, because we have been there in Europe and we have come back from it.

John Oller. I think it is incorrect to say that in an actual educational setting we would find texts that are general, that do not apply to a particular factual context. The example that Dan showed would be ruled out since you can't talk about reading a map of a particular region. What you would need is some general map of some general city, but isn't practical. What we want is a map of some actual place. If one accepts the theoretical argument that Dewey puts forward, there isn't any such thing as a general text. There are only texts that refer to actual contexts where the facts are determined and fixed. I am in agreement with the idea that we need to test grammar, propositional coherence, and the other things that Chris mentioned, but the only way to test them is with texts that have particular facts in real-life settings in some sense.

On the matter of authenticity and interestingness, I do think these are important. I agree with Lyle that these things are related. But the variables of authenticity and interest do not refer to good writing in the traditional sense of good writing. They don't mean well-formed prose in the sense that the English teacher would view it. What they mean is that good writing is the sort of thing that would appeal, that people would read. Now I found the task that Chris gave us about the psychiatrist to be intrinsically interesting because it has a powerful personal content; it engages the student and focuses his interest on the text. I think we can do research that will determine whether that sort of engagement can be done better by certain types of text than others. In my mind there is no question but that some texts will work better than others in terms of engaging the students. And if we focus our attention on texts that are interesting, and exclude all the boring, dull, and trite texts that sometimes creep into tests, we will be better off.

Take the example on the TSE of the item, "Describe a telephone in as much detail as you can." I don't get motivated by it. But if we change the task slightly and introduce a little disequalibrium, as in "Tell me about somebody who hates your guts" or "Tell me about somebody you can't stand to be around," you can get almost anybody motivated to talk. Authenticity is engaging the person's willingness to communicate and to use his or her compensatory strategies in doing so.

Frances Hinofotis. I think we have to be careful about affective reactions to things on the test that are negative. In fact, I would argue that we don't want very much on the test to be negative, nor do we want to cause any affective reaction by the student.

I want to agree with Charlie's observation that much real-world writing is ineloquent. In fact, I would take his statement even further and say that some of it is incomprehensible. I would say that it is acceptable to use ineloquent language on a test, provided you give it to a native speaker first and verify that he or she can understand it and answer the same questions you are asking the nonnative speaker. When we put authentic language on a test, we are sometimes asking the nonnative speaker to answer questions a native speaker couldn't handle, and I think this is unfair.

Chris Candlin. I want to comment about the distinction between authenticity of text and authenticity of task. I think it is going to be difficult to identify authentic texts because of the great variety of consumers we have. I also think it will be difficult to select tasks that will have equivalent interest levels, because we all have personal views as to what is interesting and what is not interesting. However, what we might do is devise a set of test questions involving particular academic-related tasks that people have to perform across a range of disciplines and attach these same tasks to any variety of texts, so we can design tests in which the process is specified, but the text on which the process was to be engaged was much more free-floating. Again, I think we should try to identify the tasks, at the level of cognitive operations, to which there could be applied a variety of different texts. So with TOEFL's twelve tests per year, you would change the texts but keep the processes the same.

Alison d'Anglejan-Chatillon. I have some concerns about authentic unscripted speech. In Canada, I worked on the Canadian version of "Sesame Street." We found that when we used unscripted speech, the people found it so unacceptable that they didn't want any part of it. If you are going to attempt to use more authentic speech, I think it should be scripted. If you use unscripted speech on the TOEFL, you would be asking nonnative speakers to encounter something that they have never encountered before on a test. While they may be able to adapt very quickly to this in a real context, you would be asking them to do this for the first time and under the pressure of a stressful situation.

Another concern I have about authentic speech relates to local dialects. When I was listening to the tape of unscripted speech that Dan Douglas played, I had some trouble with these dialects myself.

Karin Steinhaus. With respect to genuine texts, in the last few years in the TOEFL test development group we have been using genuine authentic texts, rather than polishing the language of the text. It has made finding suitable texts much harder. But it has also gotten us off the hook because we previously felt obligated to fix the inadequacies in the text or add something else to the text so we could get a few more questions out of it. In fixing them, we sometimes introduced inaccurate interpretations into the text. This was because the texts involved slight levels of specialization, and it turned out that operating at these levels of specialization was beyond our ability.

Francine Steiglitz. When the Committee of Examiners reads passages for the TOEFL the passages do seem to be lacking in authenticity, probably because they are so short and are out of context. However, the same passage read as part of a larger text in a biology course would not be so lacking in authenticity. When examinees arrive at the Reading Comprehension section of the test, the passages may not seem so inauthentic as when we see them outside of that context.

Protase Woodford. We are sometimes told that our tests are bland. However, in developing tests we try to make them interesting. But, in doing so we become concerned that we will create distraction. This might not be a problem for a literature teacher in a classroom, but on a test it is. For example, on the Listening Comprehension section, examinees have twelve seconds to read the options and answer the question. If the passage deals with somebody dying, there will be examinees who just had deaths in their families. We worry constantly about someone writing us that the content of the test upset them.

Joy Reid. When we talk about measuring the communicative competence of TOEFL examinees, are we refering to those who are at the admission level? Are we also talking about the examinee who is at a lower level of langauge proficiency? In terms of difficulty, is there a hierarchy of different kinds of competency that TOEFL could test? Do we know enough to overlay different kinds of communicative competence questions at different levels of language proficiency?

Richard Duran. In the analysis we did of the TOEFL (Duran et al. 1985), Michael Canale suggested that, since the ILR scale desriptions represented a hierarchical scale of general competency, it might be interesting to relate TOEFL items to the scale. This raises the issue of coming up with a simple hierarchy of competencies. If this can be done, it would also be of value to the examinee in that it would provide a practical interpretation of test performance.

Lyle Bachman. I think Joy's questions focus on whether we can talk about a uniform configuration of language proficiency for aspects of communicative competence across levels. I don't think we can. One of the things that led to the reformulation of the FSI scale to the current ILR scale is that fact that FSI examiners found that people at the same proficiency level often had different areas of competency. There were people who were good in grammar but who were very weak in dealing with a particular function. Or, there were people who were very good in being able to describe but their grammar was poor, and so forth. A few years ago Ray Clifford hypothesized that people at level 1 might be better in their pragmatic competence than in their grammatical competence. I think this is intuitively correct on the basis of my experience. We have been talking about the aspects of communicative competence as if they all exist in equal proportion and are of equal importance across a whole range of proficiency levels, and that simply is not the case. This suggests that one approach the TOEFL program should consider is that of a multilevel test.

Charles Stansfield. In regard to the ILR scale, we found that it was only applicable to describing the difficulty of a test when there was a considerable amount of context. So it isn't useful in discussing discrete-point items, because there is not enough context to judge the item. But where there is context, the level is related to the propositional complexity of the text. And propositional complexity has to do with how well the passage is written. If the passage is not well written, if the ideas do not flow, if it is not both cohesive and coherent, then the examinee has to make a lot of inferences as to the meaning. This increases the propositional load in that more propositions are invoked, but they are not made explicit. In order for the text to make them explicit, it would have to be made more coherent and cohesive. This would usually make the text longer, and a longer text is less complex than a shorter one that contains the same number of propositions. I think this suggests the possibility of scaling passages according to their propositional complexity, which is a kind of criterion-referencing that could be applied to the selection of passages for tests of different levels of proficiency.

I think that the Pre-TOEFL does a good job of addressing Lyle's suggestion that we develop tests for different levels. The Pre-TOEFL is based on genuine texts taken from the same sources as the TOEFL. But, perhaps because of the material's lower propositional complexity, we know from statistics that it is suitable for lower-level examinees. I don't know if the type of material that should go into a lower-level test would have to come from different sources. If it were selected for statistical reasons from the same sources, it would give us tests at different levels. If the texts were selected from completely different sources, then comparisons between individuals might not be possible. Still, when comparing examinees on similar tests it might be possible to make practical statements about their skills if we had a scale that related the texts to the competencies they require. So, I would see the texts coming from the same sources but invoking different levels of competency.

Harold Madsen. At our university, we have our own test for placing students at various levels, even though we do rely on TOEFL for admissions purposes, so we are very concerned about providing a sort of hierarchical arrangement of the tasks. It would seem to me to be appropriate to relate these to a minimally acceptable level of performance at the university level. I have some negative reactions to the idea that tasks have to be completely natural. We once had a teacher who could read a script so well that you couldn't understand him. So, as we move toward naturalistic text, we have to modify it in such a way that it is comprehensible. I do think the move toward naturalistic text could have a backwash in terms of instruction. If examinees say, "But we've never heard that kind of English before," the school should make sure that they do the kind of English that involves natural expression.

References

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 17). Princeton, NJ: Educational Testing Service.

Engelskirchen, A., Cottrell, E., & Oller, J. W., Jr. (1981). A study of the reliability and validity of the Ilyin Oral Interview. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), The construct validation of tests of communicative competence (pp. 83-93). Washington, DC: Teachers of English to Speakers of Other Languages.

Grice, P. (1975). Logic and conversation. In P. Cole & M. L. Morgan (Eds.), Syntax and semantics: Speech acts. New York: Academic Press.

Huckin, T., & Olsen, E. (1984). On the use of informants in LSP discourse analysis. In A. K. Pugh & J. M. Ulijn (Eds.), Reading for professional purposes: Studies in native and foreign languages (pp. 120-129). London: Heineman.

Ilyin, D. (1976). Ilyin Oral Interview. Rowley, MA: Newbury House.

Johns, T. (1981, July). Team teaching. Lecture presented at the University of Aston, United Kingdom.

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

Selinker, L. (1979). On the use of informants in discourse analysis and language for specialized purposes. International Review of Applied Linguistics, 17, 189-215.

CLOSING REMARKS

G. Richard Tucker

In one sense this is probably the most tightly focused conference that I have attended in a long time. We had papers distributed beforehand. We had oral summaries presented here of the papers, and we had oral responses to the papers that were prepared in advance and then were delivered orally here. We also had questions on the papers and responses and a discussion of each. In this final portion of the program, I would like to review the substance of our discussions over the entire conference. I want to make very brief remarks in five areas.

First, some general background remarks. I think that we are all delighted to have had the opportunity to participate in this Second TOEFL Invitational Conference. The conference grew out of discussions among the TOEFL Committee of Examiners, the Research Committee, the Policy Council, and, of course, TOEFL program staff. The purpose was to examine the testing of communicative competence and, indeed, the notion of communicative competence within the context of the TOEFL. By extention, we, of course are asking whether the current TOEFL adequately and accurately assesses an individual's ability to participate effectively in tertiary instruction. We were asked to do so with an open mind and with very few constraints, other than that at some point Russ Webster, Woody Woodford, and Charlie Stansfield would have to contend with the serious logistical constraints that affect the administration of a test to hundreds of thousands of individuals each year in a format that is rapidly scorable and reportable, and using instruments that are reliable, valid, and secure. Nevertheless, we were asked to look toward the future. My only regret about the way we have done that is that perhaps we did not force ourselves to look at alternative delivery mechanisms, but I think this was due to time constraints. That's not necessarily a fault of our discussion because we did accomplish many things, but I think it's important to note parenthetically that the Star Wars technology is here. People at the University of Nebraska, at Brigham Young University, and at the Center for Applied Linguistics are using technology for teaching and testing. At CAL we are developing computer-assisted interactive videodisc-mediated ESL testing materials with accompanying random-access audio that will be in operation in Indonesia by May 1985. Of course, the advent of such advanced technology does not offer solutions to the problems of administering 400,000 TOEFL tests in various places around the world on the same date and under the same conditions. However, in terms of future deliberations, we want to make sure that we have questions concerning alternate delivery mechanisms foremost in our minds. I know ETS has these concerns in mind, and I believe that the TOEFL program has them in mind as well. Perhaps at some future date an ad hoc committee might look specifically at this, or specialists might work with Russ, Charlie, Woody, and others to think through these issues. Perhaps you are already doing that.

We were reminded in preparing for this invitational conference that while we were examining the TOEFL, the issues that would be raised might also be relevant to the Pre-TOEFL, to the SLEP (Secondary Level English Proficiency test), to the TSE, and to SPEAK. They might also be relevant to the TOEIC (Test of English for International Communication) and the Pre-TOEIC, although these tests are not part of the TOEFL program battery of tests. We might also want to talk about issues of a broader nature that would also be relevant to TOEFL.

On behalf of all the participants, I would like to note that we appreciate the eagerness and the openness with which the TOEFL staff convened and arranged this conference--a conference that involved not only the Research Committee and the Committee of Examiners, but also people from the outside. This does not always happen in undertakings of this type. I think it is indicative of the confidence, maturity, and sincerity with which the TOEFL program staff approach their work that they drew into these deliberations people whom the members of the Research and Examiners committees jointly considered the best individuals the field has to offer.

I want to make a couple of comments about the nature of our discussions. We raised explicitly the issue of whether one needs to achieve some threshhold level of proficiency to be able to participate effectively in an American academic setting. We didn't come to any specific answers as to what this threshhold level might be. We raised the question of whether we can, or whether we need to specify in greater detail, what are the language demands placed on foreign students in an American academic setting. Does the present TOEFL accurately and validly assess whether individual X will be able to participate effectively? Diane alluded to this in her remarks, as did others.

We talked a good deal about the uses of the TOEFL for admissions purposes, and we talked about the guidance that perhaps needs to be offered to admissions officers, counselors, and foreign student advisers. We talked about the consequences of making the wrong admissions decisions. We talked about the economic consequences, and we talked about the opportunity costs. We talked about the latter as an important social and moral issue, and I think we felt very strongly about that. We noted that the fact that someone does not score high enough on TOEFL to be admitted does not necessarily mean that he or she is not prepared for academic work because there are idiosyncratic examples of people who are successful. But, on the other hand it may. We talked about how we would approach the research agenda that Kathy shared with us, but obviously additional thought is needed in that particular arena. We discussed whether the present form of the TOEFL with some slight modifications would do a better job. Interestingly, we didn't say, "Let's throw out Section 2 entirely" or, "Let's make it into a five-section test and add elements x and y." What we were talking about for the most part was fine-tuning involving adjustments and attention to detail, with the exception of the addition of a writing section, which Charlie informs us is already underway. We raised the question of whether the form, the content, and the underlying theoretical rationale on which the TOEFL was based are consonant with current theory. Although we explicitly raised those questions, much of that discussion was a dichotomy that I'll make at the risk of oversimplifying. The dichotomy is, should theory drive test development and research or should practical problems and limitations drive test development and research? I realize that these are not polar opposites, but a good deal of our discussion focused on this question, how one might approach it, what one's philosophy should be, and so on.

There were a number of general concerns that were raised. Frances, Sandy, Chris, John, and others were concerned with the authenticity of language. Charlie told us about specific steps that have been taken and guidelines that have been given to test development staff to make sure that this issue will be a salient one in the construction of future forms of the test. We also talked about lack of authenticity of text to task, and we had a number of concerns about that. I'm sure that is a salient theme that will come back to the Committee of Examiners. Chris talked about problems of sampling and predictability, about establishing scales of appropriateness of acceptability, and of the exercise of creative capacity. We never really came to grips with what we will do with those issues.

We talked about how to use context, what concerns one must have with context. We had concerns with about discourse domain, that is, whether the entire academic environment constitutes a domain or whether the test itself constitutes a domain. We had concerns about the dimensions of intelligibility and acceptability. We talked about the need to express and to negotiate meaning in an American academic context. We talked about the fact that we must measure what candidates need in order to be able to participate effectively in an American academic context, and we got into a series of discussions with regard to intelligibility versus acceptability versus accuracy, with Carlos reminding us that there are situations in which errors of accuracy can be stigmatizing and the consequences quite devastating. Obviously, there is additional work that needs to be done there.

We had quite an extensive discussion concerning interest level and what guidelines there might be for item writers regarding this. We talked about the need to provide very detailed specifications for items and to identify for test development staff all relevant information in order to help them develop better items. I'm sure that many of these specifications will come out during the discussions that the Committee of Examiners will have shortly. We were concerned with the appropriateness of the test development model that we might choose and whether we should refer to criterion- or norm-referenced testing. Frances felt that it would be inappropriate for us to expect nonnative speakers, i.e., foreign learners of English, to be able to carry out tasks or to demonstrate performance that we would not expect native speakers to be able to do or that was not appropriate for native speakers.

We were obviously concerned with IRT equating. The question was raised (and I deliberated asked us not to spend a lot of time on it because we didn't possess all of the conceptual or programmatic expertise to be able to do this) of whether IRT equating was necessary or whether, in fact, the dependence upon IRT equating was unnecessarily restrictive. Lyle wanted us to consider this, and indeed I know that people here will be considering it that and Marilyn Hicks and others will be working on issues concerning IRT and TOEFL.

With respect to research, I think there was some degree of endorsement of the seven priority areas that comprised the essence of the long-term research agenda (Stansfield, 1983) that was established by the TOEFL Research Committee. You will remember that the long-term research agenda stated that we need to be able to describe better the language tasks or the demands that are made on students in an American academic context. What is it that they are expected to be able to do in speaking, listening, reading, and writing? The committee wanted to better understand the dimensions of communicative competence and to explore them. It wanted to understand better the repercussions of, and the existence of, individual and group differences. It was concerned with test validity, with equating procedures, with the development of new instrumentation, and with technology. In our discussions here during the past two days, we came back at one time or another to almost every one of those issues and made comments about their appropriateness, necessity, and relevance, and I think it's instructive that the Research Committee (according to what Henry told us) has projects underway right now in many of these areas and has precis and proposals that are pending approval in others. There were one or two specific items that came up that clamored for research attention. While they certainly fit within the long-term research agenda, they deserve mention as separate items.

The first is, what are the compensatory strategies that nonnative speakers employ? What repertoire of strategies do they have at their disposal? Dan and other people talked about that, and I want to reiterate that there is some food for thought here. John reminded us that we should be concerned with examinee performance and with error analysis and that perhaps the kinds of things that Jack Richards was writing about in 1971 are still relevant today. Chris reminded us about the need to consider the learner's interpretative capacity. He noted that people vary in their ability to communicate without trespassing on others' domains, and that somehow we have to assess the learner against a criterion of interpretative adequacy. We didn't get very far in specifying how that might be done, but again, within the domain of the TOEFL research program, this is certainly something that one might well look at. We talked a bit about test types and about item revision, formats, and innovations. I think we are all familiar enough with cloze, dictation, copy testing, ESP-type testing, and the Test of Written English. The last remark that might be made repeats an observation by Diane that evoked some comment by Lyle and by others. Early in the conference, Diane noted that the TOEFL works psychometrically but wondered whether we would be able to identify or to measure all the components of language proficiency or of communicative compentence with fine detail, and whether this effort would be useful or necessary. We concluded that perhaps not all traits or competencies are equally relevant or important for the TOEFL clientele and perhaps they are not even all measureable. Along with that we agreed with the last comment that Diane offered, that we must be very sure of the philosophy that guides test development. Our preference is to have examinees demonstrate what they do know rather than what they don't know. We should stage the test-taking event in such a way that the candidate shows us what he or she can do under the circumstances. In other words, we should bias for the best that the student can do.

I would like to thank the speakers, both the presenters and the reacters, for the care that they took in preparing their oral and written presentations, for adhering to strict time constraints, and for sharing their thoughts with us in a succinct, cogent, and timely fashion. We thank the TOEFL program staff, the organizers of the conference, for the care and diligence that they have once again shown in putting together this event. Lastly, we want to wish good luck to Russ Webster and his staff as they try to digest and implement all that has been said here during the past two days.

References

Richards, J. C. (1971). A noncontrastive approach to error analysis. English Language Teaching, 25, 204-219.

Stansfield, C.W. (1983). Review of recent TOEFL research and directions for the future. Unpublished document. Princeton, NJ: Educational Testing Service.

73599-027715 • RR23C0.2 • Printed in U.S.A.
I.N. 275581