DOCUMENT RESUME

ED 296 612 FL 017 539

AUTHOR Lowe, Pardee, Jr., Ed.; Stansfield, Charles W., Ed.

TITLE Second Language Proficiency Assessment: Current Issues. Language in Education: Theory and Practice, No. 70.

INSTITUTION ERIC Clearinghouse on Languages and Linguistics, Washington, D.C.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO ISBN-0-13-798398-0
PUB DATE 88
CONTRACT 400-86-0019
NOTE 207p.
AVAILABLE FROM Prentice-Hall, Inc., Book Distribution Center, Route 59 at Brook Hill Dr., West Nyack, NY 10994.
PUB TYPE Information Analyses - ERIC Information Analysis Products (071) -- Information Analyses (070)

EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS Educational History; Evaluation Criteria; *Language Proficiency; *Language Tests; *Reading Skills; Research Needs; *Second Languages; Test Theory; *Uncommonly Taught Languages; *Writing Skills

ABSTRACT

A collection of essays on current issues in the field of second language proficiency assessment includes: "The Unassimilated History" (Pardee Lowe, Jr.), which chronicles the development of proficiency testing; "A Research Agenda" (John L. D. Clark and John Lett), a discussion of research considerations and needs in proficiency testing; "Issues Concerning the Less Commonly Taught Languages" (Irene Thompson, Richard T. Thompson, and David Hiple), which examines the relevance and appropriateness of proficiency testing theory and practice for less commonly taught languages; "Issues in Reading Proficiency Assessment", including "A Framework for Discussion" (Jim Child) and "Interpretations and Misinterpretations" (June K. Phillips), discussions of proficiency testing in the government and academic contexts; and "Issues in Writing Proficiency Assessment," including "The Government Scale" (Martha Herzog) and "The Academic Context" (Anne Katz), which look at an unexplored area in proficiency testing. (MSE)

Charles W. Stansfield, Editors

SECOND LANGUAGE PROFICIENCY ASSESSMENT

CURRENT ISSUES

Language in Education:
Theory and Practice

SECOND LANGUAGE PROFICIENCY ASSESSMENT: CURRENT ISSUES

Pardee Lowe, Jr.
and Charles W. Stansfield,
Editors

A publication of CAL Center for Applied Linguistics

Prepared by the ERIC Clearinghouse on Languages and Linguistics

PRENTICE HALL REGENTS Englewood Cliffs, New Jersey 07632

Library of Congress Cataloging-in-Publication Data

Second language proficiency assessment : current issues / Pamela Lee...
[et al.] and Charles W. Stansfield, editors ; prepared by the
Clearinghouse on Languages and Linguistics.
p. cm. -- (Language in education ; 70)
"A publication of Center for Applied Linguistics."
Bibliography: p.
ISBN 0-13-798398-0
1. Language and languages--Ability testing. 2. Language and
languages--Study and teaching--United States. I. Lee, Pamela,
1936- II. Stansfield, Charles W. III. ERIC Clearinghouse on
Languages and Linguistics. IV. Series
P53.4.S45 1988
418'.0076--dc19 88-18029
CIP

LANGUAGE IN EDUCATION: Theory and Practice 70

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education, under contract no. 400-86-0019. The opinions expressed in this report do not necessarily reflect the positions or policies of OERI or ED.

Production supervision: Arthur Maisel
Cover design: Karen Stephens
Manufacturing buyer: Art Michalez

Published 1988 by Prentice-Hall, Inc.
A Division of Simon & Schuster
Englewood Cliffs, New Jersey 07632

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN 0-13-798398-0

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Dehli*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

Language in Education: Theory and Practice

ERIC (Educational Resources Information Center) is a nationwide network of information centers, each responsible for a given educational level or field of study. ERIC is supported by the Office of Educational Research and Improvement of the U.S. Department of Education. The basic objective of ERIC is to make current developments in educational research, instruction, and personnel preparation readily accessible to educators and members of related professions.

ERIC/CLL. The ERIC Clearinghouse on Languages and Linguistics (ERIC/CLL), one of the specialized clearinghouses in the ERIC system, is operated by the Center for Applied Linguistics (CAL). ERIC/CLL is specifically responsible for the collection and dissemination of information on research in languages and linguistics and its application to language teaching and learning.

LANGUAGE IN EDUCATION: THEORY AND PRACTICE. In addition to processing information, ERIC/CLL is involved in information synthesis and analysis. The Clearinghouse commissions recognized authorities in languages and linguistics to write analyses of the current issues in their areas of specialty. The resultant documents, intended for use by educators and researchers, are published under the series title, Language in Education: Theory and Practice The series includes practical guides for classroom teachers and extensive state-of-the-art papers.

This publication may be purchased directly from Prentice-Hall, Inc., Book Distribution Center, Route 59 at Brook Hill Dr., West Nyack, NY 10994, telephone (201) 767-5049. It also will be announced in the ERIC monthly abstract journal *Resources in Education (RIE)* and will be available from the ERIC Document Reproduction Service, Computer Microfilm International Corp., 3900 Wheeler Ave., Alexandria, VA 22304. See *RIE* for ordering information and ED number.

For further information on the ERIC system, ERIC/CLL, and CAL/Clearinghouse publications, write to ERIC Clearinghouse on Languages and Linguistics, Center for Applied Linguistics, 1118 22nd St. NW, Washington, DC 20037.

Gina Doggett, Editor, Language in Education

Acknowledgments

The editors wish to acknowledge the contributions of several individuals who helped us carry out the many tasks required to bring a volume such as this to fruition. The outside reviewers, Anthony A. Ciccone, Claire J. Kramsch, and Sally Sieloff Magnan, provided excellent direction. In addition, each chapter author reviewed and critiqued at least one other chapter, and patiently responded to many suggestions, a process that resulted in multiple drafts of each chapter. Gina Doggett, the Language in Education series editor, provided general guidance and assisted us with the editing of each chapter. Jeanne Rennie and Dorry Kenyon also provided many useful editorial suggestions. Finally, we would like to thank the Office of Educational Research and Improvement of the U.S. Department of Education, through which this publication was funded.

P.L.
C.W.S.
December 1987

Contents

Introduction 1
Pardee Lowe, Jr., Charles W. Stansfield

Chapter I: The Unassimilated History 11
Pardee Lowe, Jr.

Chapter II: A Research Agenda 53
John L.D. Clark, John Lett

Chapter III: Issues Concerning the Less Commonly Taught Languages 83
Irene Thompson, Richard T. Thompson, David Hiple

Chapter IV: Issues in Reading Proficiency Assessment
Section 1: A Framework for Discussion 125
Jim Child
Section 2: Interpretations and Misinterpretations 136
June K. Phillips

Chapter V: Issues in Writing Proficiency Assessment
Section 1: The Government Scale 149
Martha Herzog
Section 2: The Academic Context 178
Anne Katz

1
Introduction

This decade has seen a surge of interest in second language proficiency assessment within the academic community. This interest follows on a long history of activity in assessing second language proficiency within the U.S. government, which began in 1956 with the initial development of the Foreign Service Institute (FSI) Oral Proficiency Rating Scale (Sollenberger, 1978). By 1973, the Interagency Language Roundtable (ILR) had assumed primary responsibility for the scale. This intergovernmental committee is comprised of representatives of government agencies that are concerned with second language teaching and testing, including the FSI, the Peace Corps, the Central Intelligency Agency, the National Security Agency, the Defense Language Institute, the Department of Education, and a number of others. Together, these organizations have considerable experience and expertise in intensive second language instruction. Drawing on this experience, the ILR Testing Committee has continued to refine the government's definitions of proficiency associated with each level and each skill on its ILR scale (Interagency Language Roundtable, 1985).

In recent years, the American Council on the Teaching of Foreign Languages (ACTFL) has developed and disseminated a derivative set of proficiency guidelines (or skill-level descriptions, as they would be called by the ILR) with the assistance of the ILR and the Educational Testing Service (ETS). This set of guidelines is designed to relate to the academic learner. This

learner typically differs from the government employee in a number of important ways. The typical student in the government setting is an adult learner in an immersion program who has both a utilitarian motive for studying a second language and the opportunity to apply classroom learning to daily job requirements in the target-language country. The academic learner, on the other hand, usually studies the foreign language as part of an effort to obtain a humanistic education. His or her exposure is typically limited to three to five hours per week of contact with the teacher in the classroom. Also, instead of completing this study of a language in one year or less of intensive instruction, the academic learner often continues to study it for two or more years. In short, students in academia learn the language in the classroom, while the government language-learning situation is characterized by learners who have an opportunity to "acquire" (Krashen, 1981) the language both inside and outside the classroom.

The fact that the government and the foreign language teaching profession typically encounter two quite divergent learner groups suggests that each has an important perspective to contribute to an understanding of second language proficiency and its assessment. For this reason, in this volume we have brought together a group of authors with experience in academia, in government, or in both.

As the influence of the ACTFL guidelines has been increasingly felt in academic circles, a number of concerns about them have arisen. Many concerns are quite legitimate, while others seem to reflect an incomplete understanding of the origin and intent of the guidelines. We hope that this volume, which attempts to both clarify the ACTFL and ILR documents and propose additional work that is needed on them, will contribute to a better understanding of this approach to language proficiency assessment and further advance the field.

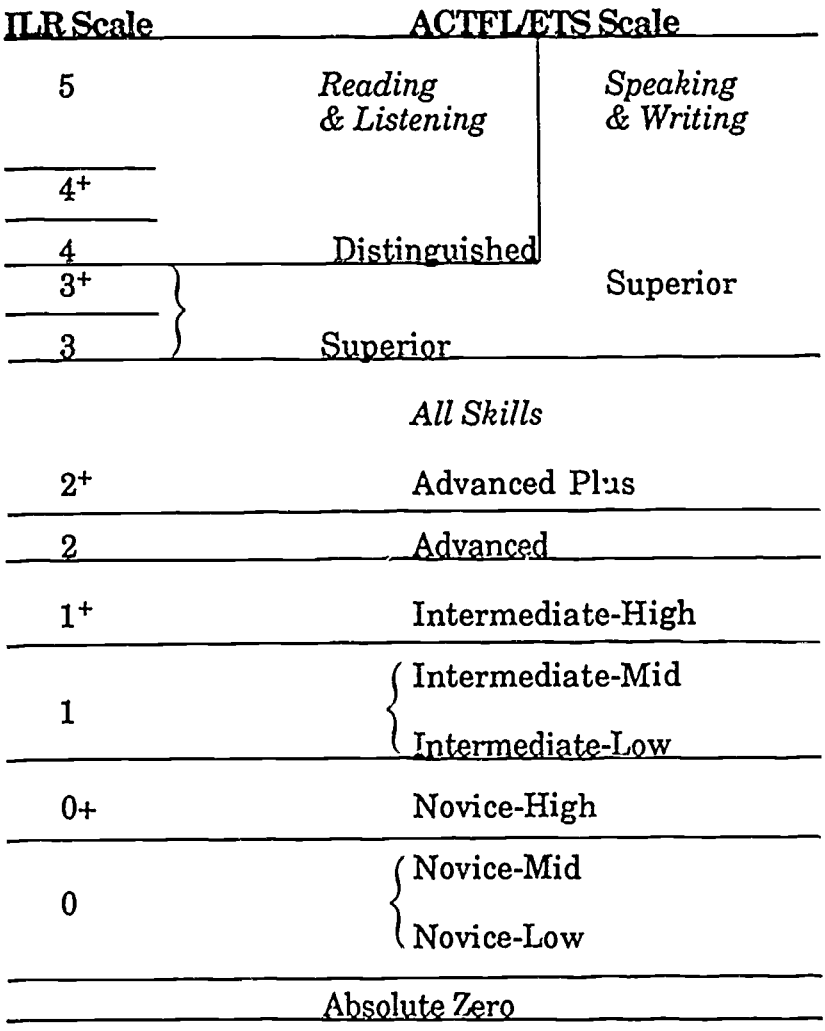
The history of today's proficiency testing movement can be traced from its foundations in one government agency (FSI) through its development with cooperating

government agencies (the ILR) to its expansion into academia through the work of ACTFL and ETS. To reflect this developmental history, the guidelines are referred to throughout this book as the AEI (ACTFL/ETS/ILR) guidelines. Although other definitions of foreign language proficiency have appeared (Bachman & Savignon, 1986; Cummins, 1984), the AEI guidelines stand apart as the most comprehensive and widely implemented definitions to date.

Differences, however, do exist between the ILR and the ACTFL/ETS scales. The ILR scale extends from 0 (for no ability to communicate effectively or understand the language) to 5 (for ability equivalent to that of a well-educated native speaker/listener/reader/writer). The scale includes pluses at Levels 0 through 4, designating performance that substantially surpasses the requirements for a given level but fails to be sustained at the next higher level, thus furnishing an 11-range scale.

The ACTFL/ETS scale, derived from the ILR scale, provides three distinctions each at the ILR 0/0$^+$ and 1/1+ levels. Thus, it is more sensitive than the ILR scale at the lower levels of proficiency. Note, however, that the ACTFL/ETS scale places all ILR levels above 3 (i.e., 3, 3$^+$, 4, 4$^+$, and 5) under an omnibus designation, Superior. The ACTFL/ETS scale thus has 9 ranges. Moreover, the ACTFL/ETS scale bears prose designations, such as ACTFL/ETS Advanced, rather than the ILR numeric designations, such as ILR Level 2. The two scales are compared on the next page.

In this volume, ten experts in second language testing draw on their experience to identify significant issues and share perceptions and concerns about possible solutions. The authors make no attempt to hide their struggle with the implications of the AEI system. Thus their writings reflect the provisional nature of available knowledge on assessing second language proficiency. The present state of flux is at least partially due to the melding of two different traditions of measurement—a top-down, holistic approach deriving from the work of the ILR, and the various uncoordinated



Relationship of ILR Scale to ACTFL/ES Scale

systems in use in academia that are traditionally bottom-up and atomistic in nature.

The volume focuses on the skills of speaking, reading, and writing. The authors of each chapter treat them separately and in mixes that best illustrate the issues they present.

Speaking, the most fully understood and studied skill in the history of proficiency assessment, forms the major focus of the first three chapters, "The Unassimilated History," "A Research Agenda," and "Issues Concerning the Less Commonly Taught Languages," but receives only passing attention in the remaining chapters. The book's concluding chapters, each containing a section on testing in the government and a section on testing in the academic sector, are devoted to the nature and assessment of reading and writing proficiency.

The first chapter deals with the causes, gaps, and misperceptions that seem to have led many to assume that the ACTFL/ETS guidelines appeared on the scene suddenly in the early 1980s. The chapter seeks to describe language proficiency assessment as it was carried out before then, both in academia and in the U.S. government. It neither attempts to apologize for nor to criticize this prior history, but maintains that this history must be understood in order to understand the AEI concept of proficiency more fully, to use it more intelligently, and to investigate it more thoroughly.

It is true that, in the decades preceding the 1980s, the relevant research conducted within the government was not disseminated to academia. However, in the 1950s, when the ILR definitions were first written, the profession may not have been ready to assimilate a type of testing that stressed functional foreign language skills, thus radically differing from academia's focus on literature and culture. In addition, there seem to be aspects of the ILR system that were never really made known outside the government, such as the realignment of tasks according to difficulty across all the skill modalities in 1978-79. Nor did manuals exist on how to administer a proficiency test. A manual on oral interview testing was not begun in earnest until 1981. To this day, there are no ILR manuals describing the testing of reading, listening, or writing. Thus, the nonassimilation of the developmental history prior to the 1982 appearance of the ACTFL/ETS Guidelines is partially understandable, but remains a significant problem to be

overcome before the nature of the AEI scales can be truly understood.

The second chapter of this volume proposes a research agenda, examining proficiency testing in light of validity and reliability concerns. The strengths of the AEI model are discussed, along with areas of needed research, future development, and possible modifications. The nature of language tested and its appropriateness and intelligibility as judged by native speakers are also treated. Suggestions are made about standardization of the interview process and optimum interview length. Attention is also devoted to the need for research into proficiency and language attrition.

The third chapter addresses the application of proficiency guidelines to the less commonly taught languages. Questions of relevance and appropriateness in theory and practice are addressed. In addition, questions of Eurocentric bias and the impact of the application of the provisional generic guidelines to languages with different typologies are traced with reference to speaking and to reading. Theoretical and practical problems in adapting the guidelines to specific languages are discussed, and finally policy issues affecting the various constituencies are raised including the role of the federal government and the language and area studies centers.

The fourth chapter's first section, on evaluating reading proficiency in the government setting, distinguishes between skill-level statements, typology of texts, and reader performance. The author proposes refinements to the reading skill-level definitions and addresses the conflict between idealized testing of reading and the practical issues that must be faced when actual performance is to be rated. The hierarchical structure of reading tasks for both native and nonnative readers is discussed, with a focus on the problem that arises when foreign language students are turned into *decoders* of advanced-level texts. An example is the Russian-language student who attempts to decipher Dostoevsky with dictionary in hand. This approach to

reading comprehension contrasts with an alternate strategy that seeks automatic comprehension of texts at a near-native reading rate.

The second section, on reading in the academic setting, discusses instructional implications of both the traditional approach and that suggested by a proficiency orientation. The chapter describes the reading process and reader strategies. It examines how the higher-level nonnative reader pulls reading strategies as well as native- and target-language linguistic ability together to read mid- and higher-level texts of a general nature.

The last chapter addresses a heretofore unexplored area of proficiency testing, writing. The first section, concerning the testing of writing proficiency in the government setting, treats the history and use of the ILR writing scale. Improvements to the scale are proposed and areas of research suggested.

The second section, concerning writing in an academic setting, examines the concept of the major writing assignment, as well as the differences between testing the writing of native English users and the testing of writing proficiency in a second or foreign language. The role of the guidelines in assessing foreign language writing competence concludes the chapter.

For reasons of both space and development, it has been impossible to include three topics related to the present discussion: listening comprehension, culture, and translation. Noting that listening tasks can be categorized into participative listening and nonparticipative listening (Phillips & Omaggio, 1984; Valdman, 1987), let it be said here that participative listening may not be as readily separated from speaking as the ACTFL/ETS Guidelines and the ILR Skill Level Descriptions imply. As central as culture is to the field of foreign language education, cultural skills differ significantly from the skills described in other guidelines, as is substantiated by the developers of the ACTFL Proficiency Culture Guidelines (see Hiple, 1987; Galloway, 1987). Thus the topic of culture lies outside the domain of the present work. Translation, too, as a bicultural,

bilingual, and multiple-skill undertaking, presents issues beyond the scope of this volume.

Lastly, because this book focuses on proficiency assessment, space permits only passing comment on the implications of proficiency for curriculum design and classroom methodology. A full treatment of its effects on curriculum is properly the subject of another volume.

A driving force behind proficiency assessment has been the search for a common metric (Educational Testing Service, 1981) of a language user's ability regardless of the text taught, instructional method used, or site of instruction. A proficiency score is thus a characterization of how well a person can function in the language in question: how well he or she speaks, writes, reads, or understands the language. Although the need for a national metric is open to debate, the establishment of a common measure will be possible only when we can adequately and accurately assess the level of skills language users possess. We hope this volume will contribute to an understanding of this measurement process by both classroom foreign language teachers and second language testing professionals.

—Pardee Lowe, Jr., Office of Training and Education, Central Intelligence Agency; and Charles W. Stansfield, Director, ERIC Clearinghouse on Languages and Linguistics

NOTE

1. The ILR is a consortium of government agencies with a need to hire, train, test, and use employees whose jobs require skills in a foreign language. The ILR meets monthly in plenary session (except in summer) and regularly in committees, including management, curriculum and research, computer-based training, and testing committees. The testing committee prepared the skill-level descriptions referred to in this volume.

References

Bachman, L.F., & Savignon, S.J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70, 291-97.*

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), *Language proficiency and academic achievement.* Clevedon/Avon, England: Multilingual Matters.

Educational Testing Service. (1981). *A common metric for language proficiency* (Final Report for Department of Education Grant No. G008001739). Princeton, NJ: Author.

Galloway, V. (1987). From defining to developing proficiency: A look at the decisions. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations, and concepts.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Hiple, D.V. (1987). A progress report on ACTFL Proficiency Guidelines, 1982-1986. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations, and concepts.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Interagency Language Roundtable. (1985). *Language skill level descriptions* (Internal Document). Washington, DC: Author. Also available in Appendix E of R.P. Duran, M. Canale, J. Penfield, C.W. Stansfield, & J.E. Liskin-Gasparro (1985). *TOEFL*

from a communicative viewpoint on language proficiency: A working paper (TOEFL Research Report 17). Princeton, NJ: Educational Testing Service.

Krashen, S.D. (1981). *Second language acquisition and second language learning.* Oxford, England: Pergamon Press.

Phillips, J.K., & Omaggio, A.C. (Eds.). (1984). [Special Issue on the 1983 ACTFL Symposium on Receptive Language Skills]. *Foreign Language Annals, 17*(4).

Sollenberber, H.E. (1978). Development and current use of the FSI Oral Interview Test. In J.L.D. Clark (Ed.), *Direct testing of speaking proficiency: Theory and application.* Princeton, NJ: Educational Testing Service.

Valdman, A. (Ed.) (1987). *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency* (March 1987). Bloomington, IN: Indiana University.

I
The Unassimilated History

by Pardee Lowe, Jr.,
Office of Training and Education, CIA

As the proficiency movement[1] matures, there seem to be as many opinions on the nature of proficiency and its merits as there are participants in the debate. Some confusion surrounds the genesis of the movement, which is thought by many to have begun in 1982, when the ACTFL Provisional Proficiency Guidelines were published. These are often called the ACTFL/ETS guidelines because the American Council on the Teaching of Foreign Languages (ACTFL) and the Educational Testing Service (ETS) cooperated in their production. When the guidelines were revised in 1986, the qualifier "provisional" was dropped, and they became the ACTFL Proficiency Guidelines.

Proficiency testing procedures[2] were used by the government for years before the guidelines were developed for use in academia (Jones, 1975; Liskin-Gasparro, 1984a; Lowe, 1985b; Sollenberger, 1978; Wilds, 1975). However, until now, the early stages of the guidelines' history seem to have been unassimilated by the second language education community. It is the task of this chapter to record this unassimiliated history in the hope of clearing up the misunderstandings that consistently arise in the current literature and at conferences where proficiency is discussed. However, this chapter is not intended for those who are new to the proficiency

not intended for those who are new to the proficiency concept. It necessarily begins in *medias res,* specifically addressing those already generally familiar with proficiency testing and struggling with the issues it presents—issues that delving into its prior history should help to clarify.

Defining AEI Proficiency

The designation AEI is a ready clue to the guidelines' origins. To be historically correct, the acronym should be ILR/ETS/ACTFL, because after the State Department's Foreign Service Institute (FSI) originated the proficiency assessment system, the Interagency Language Roundtable (ILR) refined it. It was first used outside the government by ETS to rate the language proficiency of Peace Corps volunteers. ETS later, through the Common Metric Project in 1979, examined the system's suitability for use in academia (ETS, 1981). Finally, ACTFL adopted ETS' recommendations and produced the guidelines much as they exist today.

The ACTFL/ETS scale was purposely derived from and designed to be commensurate with the ILR scale. As a result, the designation AEI expressly refers to the many facets the two systems share. To reflect this joint history, an operational definition outlining a common core might read:

> AEI proficiency equals achievement (ILR functions, content, accuracy) plus functional evidence of internalized strategies for creativity expressed in a single global rating of general language ability over a wide range of functions and topics at any given ILR level. (Lowe 1985b, 1986.)[3]

Achievement is used here as it is traditionally understood in the academic classroom as well as in the

government sense of mastery of ILR functions, content, and accuracy when these are memorized but not used creatively in a consistent and sustained manner. *Functional* refers to obvious, usable skills; that is, those that are proven by use, not merely implied. For example, the presence of a conditional does not prove consistent and sustained ability to hypothesize.

The need for a fuller definition of proficiency is clear, although simple definitions exist. Shulz (1986) mentions "the ability to send and receive different messages in different real life situations" (p. 374). Larson and Jones (1984) summarize definitions of proficiency before 1984 as "the ability to communicate accurately." Lowe (1986, p. 392) also offers the following short characterization: "doing things through language." Convenient as these short characterizations of proficiency may be, they can also be misleading.

Larson and Jones' review of the literature (1984) suggests that everyone knows what "proficiency" is. Unfortunately, any two observers who have not received training in AEI proficiency assessment, when asked to assess the "proficiency" of a performance, will select different aspects to rate; their definitions of proficiency would differ accordingly.

Webster's New International Dictionary suggests sources for such differing interpretations of proficiency by presenting a continuum of definitions ranging from indications of partial control (progression, advancement) to proof of expert control. While the dictionary stresses expertness, popular usage in foreign language circles tends to stress advancement. But this leaves several questions unanswered:

1. What constitutes advancement?
2. How is advancement quantified?
3. How much advancement is involved?
4. Is that advancement significant?

The existence of an abundance of definitions of proficiency in the foreign language profession, from

any advancement deemed significant by the observer to (near) expertness, is demonstrated by Larson and Jones' (1984) survey of proficiency definitions:

> Certainly germane to the issue of proficiency is an understanding of what is to be meant by "proficiency." Interpretations of the term range from the ability simply to use properly the phonology and structural devices of the language . . . to complete mastery of all the components of communication. . . . Others . . . have suggested that language proficiency depends on linguistic functions, situational contexts, and personal needs. (p. 114)

But this range of interpretations proves too wide to be useful; one person's "proficiency" might be another's "achievement." Clifford (as quoted in Buck, 1984, p. 311) segments this continuum into three parts:

> An achievement test checks whether language *exists;* a performance test checks whether language is *used*; a proficiency test checks whether language is *used creatively*.

To the third category might be added "and accurately."

A more complete definition of proficiency than those mentioned by Schulz or Larson and Jones is needed because it is necessary to specify exactly how to determine the presence and extent of proficiency. The present operational definition specifies that performances be judged against ILR functions, content, and accuracy (see the appropriate trisection).[4] In considering the nature of other conceptualizations of proficiency such as the one posited by Bachman and Savignon (1986), the AEI division into ILR functions, content, and accuracy has more than passing significance. This division suggests that Bachman and Savignon's concept of Communicative Language Proficiency (CLP), renamed Communicative Language Ability (CLA), and AEI proficiency may prove incompatible. The position of

accuracy within CLA remains unclear; nor is it clear that CLA is not an atomistic, bottom-up view of language ability rather than a holistic, top-down view like that of AEI proficiency (Bachman, personal communication). These possible differences have considerable significance for any research agenda.

AEI History and Possible Hindrances to Its Assimilation

A historical context is important to the understanding of a concept's true nature. This is especially so in the case of AEI proficiency, because disregarding the past also poses a risk of altering the underlying definitions and testing procedures. As a result, the original purpose—the assessment of foreign language users' abilities according to a consistent scale—may be defeated. Although the ILR definitions have addressed this purpose satisfactorily for more than 30 years, only recently has essential information about the AEI scales begun to emege in a systematic and integrated fashion. This chapter joins in the effort to document the historical context of the proficiency movement.

The following discussion rests on 14 years' experience in administering and certifying interviews of oral language skills and reading comprehension; training oral interviewers and oral interview trainers; training test designers and item writers; refining the ILR Skill Level Descriptions; and developing and revising the ACTFL Proficiency Guidelines.

In-House ILR Research

To researchers, a major drawback of the AEI scales has been the lack of research data. Studies conducted to

date afford little understanding of the rationale behind the government's research decisions (Lowe, 1985b). The limited amount of early research compels the formulation of a broadly based research agenda (Bachman & Clark, 1987; Clark & Lett, this volume). Academia's initial lack of interest in the system was another reason for the small number of studies available.

Proficiency evaluation was developed when academia focused heavily on literature. The government did not set out to create a radically new testing system. Linguists at FSI would have gladly adopted or modified a system for ascertaining oral proficiency if one existed; the fact was that none did. Francis Cartier, former chief of evaluation at the Defense Language Institute (DLI) in Monterey, Calif., enunciated the dictum by which most ILR linguist-managers must abide: "Theory has no deadline! I must have a Spanish reading proficiency test in place by next Tuesday" (comment made at ILR pre-Georgetown University Round Table [GURT] symposium, March 1980). And so the government set out to design and implement its own test.

Not generally known is that the requisite expertise existed within the government, and that it derived from academia. Claudia Wilds, a primary designer of oral proficiency assessment procedures and later head of testing at the FSI School of Language Studies, was a former student of John B. Carroll, then at Harvard University. Under Wilds' influence, FSI drew on the latest psychometric theory of the early 1950s. The ILR system adopted two characteristics of this theory: criterion-referenced testing and Osgood's semantic differential (Osgood, Suci, & Tannenbaum, 1957). Both were pioneering applications in foreign language testing.

Most government research was conducted in-house and not published, a fact that has led some to believe that little research was conducted. In retrospect, this practice was unfortunate. Still, the single early publication on proficiency testing (Rice, 1959) aroused little notice in academia, with its radically different concerns.

Today it is difficult to determine exactly which aspects of proficiency were scrutinized psychometrically. Rice (1959) indicates that efforts to identify the factors contributing to the global score were undertaken and regression equations were calculated to ascertain the extent of their contribution at Level 3/Superior. Except for Rice's article, however, little else was reported from 1954 to 1978. A study of interagency interrater reliability was conducted by FSI and the Central Intelligence Agency staff in 1973-74 (Wilds & Jones, 1974). (Interrater reliability is a measure of the degree of agreement between the separate global scores when two raters rate the same performance independently.) Jones, then at CIA, who designed and executed the study with Wilds at FSI, reported that correlations between raters at the two agencies in each of three languages—French, German, and Spanish—surpassed .89, demonstrating that suitable reliability could be attained.[5] The full study is still unpublished.

In 1974, Jones and Spolsky organized an extraordinary session at the pre-GURT symposium devoted to proficiency testing. Testing experts from academia and the government participated in a discussion of the role of proficiency in U.S. language testing and training. The symposium inaugurated an annual symposium series that sporadically produces proceedings (Clark 1978; Frith, 1980; Jones & Spolsky, 1975). Even then, however, interest remained confined to a small group, primarily within the government.

Interest in the ILR's proficiency system emerged in academia in the late 1970s—by which time language requirements at universities had been eliminated or reduced drastically. A telling criticism in discussions of educational reform was the statement: "I studied _____(name your language!) for four years and can't order a cup of coffee!" Americans are a practical breed, and this utilitarian criticism hit home. Other factors came into play, such as the growth of overseas language training programs and a greater awareness among students of their lack of language competence.

In addition, U.S. foreign language educators became more familiar with European functional-notional syllabi. Concurrently, dissatisfaction grew within the foreign language community with the *status quo*. Putting aside structural linguistics, linguists were beginning to branch out into conversational analysis, discourse analysis, sociolinguistics, and even psycholinguistics. All of these factors put pressure on foreign language departments. Once again a need emerged, this time in the context of public education, for the teaching and testing of functional foreign language skills—precisely the situation faced by the State Department in the early 1950s.

The need in academia was met by a happy series of events:

1. testing workshops conducted by James Frith, dean of FSI's School of Languages, and Marianne Adams, FSI's head of testing;
2. the ETS Common Yardstick project (1981); and
3. the ACTFL/ETS Guidelines project.

All three efforts had ILR support and led to a derivative, commensurate scale for academia. These events are well-documented by Frith (1979), Liskin-Gasparro (1984a), Murphy and Jiménez (1984), Hiple (1984, 1986), and Lowe (1985b). In 1974, the government began assembling a manual on the oral proficiency interview, which was not completed until 1985 (Lowe, 1985a).

In the mid-1980s, many observers, some of whom later espoused or criticized the ACTFL/ETS Guidelines, maintained that the ILR conducted little research, rarely published it, and failed to systematize and integrate the government's experience and findings into the mainstream of American foreign language test and curriculum development (Shohamy, 1987). These critics were generally accurate in their assessment, for the period extending from the 1950s until 1974. From 1974 to 1978, a small group of academicians became aware of the ILR system, but little use was made of either the awareness or the system. From 1978 on, however, the

exchange grew; ACTFL and ETS became observers at meetings of the ILR and participated in the revision of the ILR Skill Level Descriptions, which led to the appearance of the provisional ACTFL/ETS Guidelines in 1982 (Hiple, 1987). Starting in 1974 with the Jones and Spolsky symposium and gathering momentum from 1978 to the present, the ILR has conducted considerable research on the nature of AEI proficiency (Lowe, 1985b).

In an academic setting, this research would have been carried out early in the system's existence. In the government context, a practical approach was an immediate necessity. Fortunately, the wedding of academia's and the government's current interests has culminated in a proposed common research agenda (Bachman & Clark, 1987; Clark & Lett, this volume).

Any discussion of the significant issues in AEI proficiency should build on the fullest possible understanding of its nature. While an operational definition has been proposed, a discussion of difficulties in wording the definitions, a fuller characterization of the AEI framework, and a review of the framework's applications and testing procedures should aid understanding the chapters that follow. The problems cited here have proven to be stumbling blocks to many seeking to assimilate the system.

Words Versus Experience

A crucial misperception of the AEI scales assumes that their wording reflects the system. Every deviation in wording between versions may be thought to imply notable differences. This is not usually the case. The drafters of the ACTFL/ETS Guidelines and the ILR Skill Level Descriptions took exceptional care to ensure commensurability.[6] Different wording usually signals *no* difference between the two systems, but rather an alternate way of expressing the same aspect of proficiency evaluation.

The different wording results in part from the fact that ILR base ranges were split into ACTFL/ETS subranges; for example, ILR Level 1 was subdivided into ACTFL/ETS Intermediate-Low and Intermediate-Mid (see figure in the Introduction to this volume). ILR raters had used arrows, assigning an upward arrow for a strong performance, and a downward arrow for a weak performance. The ACTFL/ETS distinctions of Low and High performance at ILR Levels 0 and 1 were derived from this procedure of assigning upward and downward arrows (Lowe, 1980).

Problems arose when ILR experience was distilled into words for the first time. Behind differences in wording lurks the greater challenge of describing a sophisticated system verbally. A central problem is that a single document is expected to speak to three audiences: examinees, examiners, and administrators. Perhaps the moment is ripe, as Clark has contended, to produce several documents, each with a distinct goal (Clark, 1987a). When the interagency handbook on oral proficiency testing reached 500 pages, it proved so unwieldy that it had to be streamlined (Lowe, 1985a). Yet many questions still arise that such a handbook cannot address. Frankly, it may never encompass all possible situations. Thus, AEI proficiency should be viewed more as a dynamic constitution than as a Napoleonic Code.[7] Perhaps the best access to the system is offered by firsthand experience: observing tests, being trained as a tester, or being trained as a tester trainer. Indeed, the definitions cannot replace hands-on exposure. This is the case for speaking and reading (Lange & Lowe, 1987), and so may be true of all skill modalities.

Wording problems, such as *overlayering*, arose for historical reasons; problems also arose in *shifts in focus* from level to level, such as describing plus-level performances and in describing *1980s tasks in 1950s language*. These problems in wording have contributed to widespread inability to internalize the system.

Overlayering

Overlayering occurs when statements describing *learner* behaviors are superimposed on earlier statements describing *user* behaviors.[8] After a series of revisions, the scale's final form was fixed at FSI in the mid-1950s. A needs analysis of jobs at home and abroad produced a set of hierarchical categories, called *levels*, into which language performances by government language users could be sorted. Initially, the needs analysis focused on *users* and their performances in the field. When the assessment system was field-tested later at FSI, language *learners* comprised its [illegible]or population.

The confusion has been exacerbated by academia's burgeoning interest in proficiency-oriented curricula, emphasizing learners, precisely because the definitions were never intended for this population. Yet the challenge of adapting proficiency guidelines for use in the classroom continues to fascinate the profession (Liskin-Gasparro, 1984b). Moreover, the statements about ILR learner outcomes were developed in the context of the government's *intensive* language learning programs.

To some, this fact suggests that meaningful AEI levels are unattainable in academia in programs for nonmajors (Schulz, 1986). This will prove untrue for two reasons: First, AEI levels do not refer to the speaking skill modality alone—an early misunderstanding. Significant functional ability can be attained in listening, reading, and perhaps writing. Secondly, unlike intensive government training, with its constant exposure to the target language, regular classroom instruction in academia provides psychological absorption time—time for the material to "sink in." This may prove important, especially in the initial stages of foreign language instruction, and its possible significance bears investigation.

Thus AEI testing should not remain unattempted for fear that students will not demonstrate satisfactory

gains. While traditional instruction may result in meager gains in speaking skills, gains in other skills may be somewhat greater. With proficiency-oriented instruction, gains may be greater still.

Shifting Focus in the Plus-Level Descriptions

A *shift in focus* is a move from stressing one view to stressing another within or across definitions. With regard to the AEI scales, this means that a definition may emphasize the user at one point and the learner at another, usually within the same level. Less obvious are shifts in focus between levels. In the following discussion, the plus-level descriptions are examined.[9]

Writing plus-level descriptions poses particular problems. While base-level descriptions may be written in a more positive vein, plus-level descriptions by nature must cite deficiencies. By definition, plus-level performances must evince many characteristics of the next higher base level, but are not sufficiently consistent or sustained to merit the higher rating. Byrnes (1987a) characterizes the mixed nature of these performances as "turbulent." Plus-level definitions must account for a wide variety of performances, difficult to capture verbally. Perhaps plus levels could be regarded as a build-up of competence, not all of which has yet been translated into performance. The description writer must choose a particular vantage point from which to compose the description, and the focus may be different from level to level, depending on the principal need as the guideline writer sees it. The 0^+ Level (ACTFL/ETS Novice) definition, for example, targets general 0^+ Level behaviors; while the Level 2^+/Advanced Plus description stresses the worst case, the "street" or "terminal" 2^+ (defined later). Recent revisions have made every effort to remove these problems.

Alignment of Task Difficulty Across Skills

The year 1978 was a watershed year in the history of proficiency evaluation, because it is when the ILR Testing Subcommittee rethought the definitions. Before 1978, the *gestalt* nature of the ILR system had not been consistently elaborated in the definitions. The scale's early *midpoint* descriptions were vying with its more pervasive threshold descriptions, causing misunderstandings about acceptable performances. Threshold systems contrast markedly with midpoint systems.

A gestalt is any clearly recognizable constellation of factors, that is, a unit or an entity that is wholly indicative of a single level and no other (Lowe, 1980). In terms of the AEI scales, a threshold is a perceptual gestalt, a figure or a constellation of factors that are clearly present to perform the function(s) in question when the major border between levels is crossed (Lowe, 1980). ILR thresholds exist at each border between a plus level and the next higher base level: $0^+/1$, $1^+/2$, $2^+/3$, $3^+/4$, $4^+/5$.

In contrast, in a midpoint system, a weak gestalt represents the lower subrange; a strong gestalt, the mid; and a still stronger gestalt, the plus subrange. Midpoint systems are ubiquitous. The color spectrum is an example, with each color's shadings of pale, normal, and deep hues. In a threshold system, on the other hand, a weak gestalt occurs at the plus-level border; a strong gestalt occurs just above the border between the plus level and the next higher base level; and a still stronger gestalt occurs toward the mid-range of the next higher level. For example, driving is a threshold system. The plus-level below might be characterized by passing a test on the rules of the road; and the next higher level by crossing a threshold when the driver can simultaneously recognize a stop sign, release the accelerator, downshift, and brake.

Past-tense narration presents a linguistic example. At the levels below the threshold, the examinee has

been gathering the linguistic bits and pieces (vocabulary, grammar, etc.) with which to perform the functions of the next higher base level. But the examinee cannot yet put all the pieces together. Use of the past is sporadic and faltering. Once the threshold has been crossed, however, the examinee can satisfactorily perform the suitable functions. At this point, there is no doubt that past-tense narration occurs in a consistent and sustained fashion. Thus, the confusion between focusing on midpoints versus focusing on thresholds in interpreting the scale was a problem in using the earlier definitions.

Further complicating a correct interpretation of the definitions is the *offset* problem; that is, the tendency for individuals to exhibit unequal levels of competence in the various skill modalities. For example, many ILR students learn to speak, listen, and read, but possess significantly lower writing facility in the target language. For students in academia, the reverse can be true, with writing skills sometimes outpacing speaking skills.

Faced with this problem, writers of ILR Skill Level Descriptions from 1978 onward felt that a certain consistency in revising the definitions was in order. There were two possible directions, neither clearly demarcated in earlier attempts. The first, the uniform approach, would have linked levels so that a typical individual rated S-1 (speaking Level 1) would be regarded as automatically having a certain ability in listening comprehension, such as L-1. This approach was rejected because, as Lowe (1985b) showed, not all speakers show such uniformity; that is, there is not always a link between ability levels. In addition, examinees in some government language training programs show a consistent tendency to acquire some skills faster than others. Even when an offset commonly occurs in a language, its magnitude may differ by examinee. In Lowe's (1985b) sample, the offset between speaking and listening in Spanish ranged from a plus level to two whole levels.

The profile approach was chosen for the level guidelines instead, making all skill modalities parallel to speaking and to one another, in tasks, difficulty, and rating standards. Consequently, an S-1's creativity in speaking was matched by a similar "creativity" in understanding, characterized by phrases in the definitions such as "able to understand sentence-length utterances" when recombined in new ways at the same level (ACTFL, 1986). This approach requires reference to user profiles, and reflects speakers' tendency to operate at different levels in different skill modalities. A possible S-1^+ profile might be: 1^+ in speaking, 2 in listening, 3 in reading, and 2 in writing. A profile reflects a foreign language user's skills more accurately because it captures variations among them.

1980s Tasks in 1950s Language

A longstanding dilemma that has only recently been addressed is that of stating 1980s tasks in 1950s language. 1980s tasks refer to recent identifications of the many tasks requiring language. 1950s language refers to the structuralist view of language, in which the AEI scales were first expressed. This is not to imply that the scales were previously incapable of rating 1980s tasks. For example, the terms *coherence* and *cohesion*, until recently, did not appear in the definitions. Yet no high-level rating (i.e., 4, 4^+, 5) would be assigned to a speaker who could not produce a coherent or cohesive sample. This is one of many examples in which the AEI scales holistically handle areas that are just beginning to be understood atomistically.

Given growing recognition of the sophistication of language, one might wonder whether a system of the scope of the AEI scales could be devised today. The task would certainly be much more complex, because since the early 1950s, infinitely more has been learned about the nature of language and language acquisition—

through transformational grammar, discourse analysis, pragmatics, and so on—than was known then, though there is obviously more to learn. The ILR system was devised when views of language were simpler and more readily captured by a few terms. For speaking, the commonly accepted factors were: pronunciation, fluency, vocabulary, and grammar; added later were sociolinguistics, culture, and tasks accomplished.[10] Similarly, factors may be identified for the other skill modalities.[11] Some believe that such terms are misleading because they may be seen as "structuralist" in origin (Bachman & Savignon, 1986). Even the term *grammar* miscommunicates to some (Higgs, 1985; Garrett, 1986). Others overemphasize vocabulary. However, properly expanded and defined through research, manuals, and workshops, these expressions usually communicate AEI proficiency to the teacher in the classroom, the administrator behind the scenes, and the foreign language user. These terms encompass the domains usually considered when describing global performance.

Today's knowledge allows for expansion beyond phrase- or sentence-level grammatical descriptions to describing discourse structure and the realization of propositions (Byrnes, 1987a). Pragmatics can be discussed, as well as the real-world constraints on what can be said (Child, 1987). If a proficiency system had to be built today, the task would be too daunting were it not for the framework inherited through the AEI scales, a simple one that permits considerable expansion.

The framework permits expansion, refinement, and a top-down view of proficiency that provides an antidote to the bottom-up activities of achievement-oriented classrooms. In sum, the proficiency system of the early 1950s was just sophisticated enough to blaze an inviting and rewarding trail, and just simple enough to allow drafters of new guidelines to maintain their bearings in the face of new and intriguing problems in evaluation.

The AEI Framework[12]

The government's need for an evaluation system to deal with oral proficiency led to another impediment to assimilating the system, the lack of an overall statement on the AEI characteristics common to all skill modalities. This section attempts to address that need.

The Outlines of the Proficiency Levels. The definition for each level outlines the target area into which characteristic behaviors fall. The definitions cite representative behaviors, content, and accuracy as needed, but make no attempt to produce an exhaustive description. For a consistent statement on the required *ILR functions, content, and accuracy* in a given skill modality, see the respective functional trisection. For details, consult the relevant manuals (ILR or ACTFL/ETS) for speaking, and the relevant literature for all skills (Educational Testing Service, 1982; Lowe 1985a, 1985b).

Stress on General Language. Language may be viewed as a continuum ranging from general language, through work-related language, to job-specific language. A major characteristic of the AEI scales is their stress on general language.

Stress on Proficiency, not Achievement. The AEI scales stress proficiency, except at the lowest levels—0, 0+/Novice-Low, Mid, and High—at which students tend to regurgitate the limited material they have learned (achievement) and possess few, if any, strategies to use the material creatively. The expected creativity must reflect the requisite quality and quantity (consistent and sustained production) for the level. "Creativity" is understood here as the ability to generate sentences and discourse that are new to the examinee, not the ability to produce immortal prose.

Quality statements are important throughout the scale. Nonetheless, at the lowest levels they play down

perfection; and indeed, even at the highest levels they assume only the "near perfection" of the well-educated native, who still occasionally lapses. Nonnatives are permitted lapses at the highest levels, provided they are native-like lapses.

Process Versus Product. It is the goal of the AEI scales to rate the examinee's facility in a given skill and not solely or even primarily on the product produced. Thus, in rating an examinee's writing proficiency, the examiner reports: "John Simpson writes French at Level 2+/Advanced Plus," and not "John Simpson's essay was a 2+." The essay obviously has to demonstrate 2+ writing skills, which means that it must be a ratable sample revealing consistent and sustained ability. It is the examinee's ability, not merely his or her performance, that receives the rating (Child, this volume).

Gestalt Nature of AEI Rating. Because of the gestalt nature of the AEI scales, rating proves most efficient and accurate when raters judge "wholes" or "near wholes" rather than bits and pieces. The system provides yet another reason to stress performance that is "consistent" and "sustained" (Lowe, 1980).

Noncompensatory Core. In the ILR scale Levels 3 through 5 are noncompensatory. This means that a core of ILR functions and certain levels of accuracy must be demonstrated for the examinee to earn the rating. Nevertheless, the lower a speaker is on the AEI scales, the more compensation is possible. An examinee's strong vocabulary, for example, may under certain conditions compensate for weak grammar at Level 2+/Advanced Plus and lower.

Even at lower levels, however, compensation is sometimes impossible. This results from the presence of a noncompensatory core; in a given sample, certain features in each language prove indispensable and must appear with the degree of quality the definitions require. For example, in West European languages, some past-tense features are part of the central core to

earn a Level 2/Advanced rating. A failure to use past tense at all, or without the requisite quality, blocks assignment of the level in question.

Criterion Shift at the Intermediate High/Advanced Border. At the 1+, 2/Intermediate-High, Advanced border or lower, the definitions state that an examinee must be able to communicate with a native used to dealing with foreigners. This often requires the listener to do most of the work—particularly at the lowest levels. The shift of responsibility materially affects the rating.

Expression of Ability in a Global Rating. The examinee's ability is expressed in a single level rating. Factors contributing to the rating have been identified, but the global score does not represent a summation of scores on the individual factors (for details, see Bachman & Savignon, 1986; Bachman & Clark, 1987; Clark, 1987a; Clifford, 1980).

Full-Range Nature of the Scale. The AEI scales cover the full range of performances from knowledge of isolated words (0, 0+/Novice levels) to ability equivalent to that of a well-educated native speaker (Level 5/the peak of Superior). The apex of the full range (ILR Level 5) is included in *both* scales because the ACTFL/ETS category of Superior comprises ILR ranges 3, 3+, 4, 4+, and 5, for which the ultimate reference point is the educated native speaker/listener/reader/writer. The term "educated native speaker" has occasioned much discussion and led to misunderstandings about the scales (Bachman & Savignon, 1986; Lowe, 1985b, 1986). Space limitations preclude a full airing here, but the confusion appears to be based partly on a failure to assimilate the wide range of behaviors the scales encompass (Lowe, 1987; Valdman, 1987).

The confusion may also be due in part to differences in terminology. ILR terms such as *school, classic, street,* and *terminal* may require explanation. School learners are formal learners who typically have a solid foundation in grammar but limited everyday

vocabulary. A classic speaker evinces a balance between solid grammar and useful vocabulary. Street learners are those who have learned a second language informally and whose everyday vocabulary—through exposure in the country of the language—is extensive (for the level in question), but whose grammar proves nonexistent, weak, or fossilized vis-à-vis the level. A terminal speaker is similar to a street learner, with language that is probably irremediable. Except for the classic speaker, these designations cannot in ILR usage apply above the Level 2+/ Advanced Plus. Because the levels above 2+ are noncompensatory, any global rating must be based on a common floor in grammar and vocabulary (see Higgs & Clifford, 1982; Lowe 1985b for details). In contrast to these few terms, academia uses a plethora of expressions (Valdman, 1987).

Since the AEI scales were first instituted, the field's understanding of sociolinguistics, culture, the relationship of standard language to dialect and regional variants has burgeoned. As the 1987 Indiana Symposium underscored, high priority should be assigned to incorporating these insights more fully into the AEI definitions (Byrnes, 1987a; 1987b).

Top-Down View of Abilities. The vantage point of the AEI scales is from the top downward; that is, the reference point is at the top of the scale, which represents the performance of a well-educated native speaker of the target language. This one aspect alone may account for the difficulty some have had in assimilating the scales. Most classroom instruction and testing assumes the opposite vantage point: from the bottom up. Bits and pieces are taught and tested, and it is hoped that by manipulating these fragments students will somehow produce larger units. In contrast, the top-down view assumes that to successfully produce units at one level, learners must have a developing sense of the larger units above that level. For example, the language of a student at the Intermediate/1 Level is often characterized by short, discrete sentences. To produce a longer,

intelligible unit, the student must first have a sense of a larger unit such as a complex or compound sentence.

The salient characteristics of top-downness include the educated native speaker at the scales' apex; the gestalt nature of AEI levels; the expanding noncompensatory core as the scale is ascended; and the full-range nature of the scale. These characteristics apply to all four skill modalities. Together they further distinguish AEI proficiency from other kinds of proficiency.

Applications and Procedures

Another factor that contributes to difficulty in internalizing the system is the wide variety of approaches to testing the nonoral skills, many of which suspiciously resemble those used in achievement testing.[13] Superficial similarities aside (e.g., in reading, multiple-choice and cloze formats), any AEI proficiency testing procedure must:

1. obtain a ratable sample, usually more extensive than that of an achievement test;
2. establish consistent and sustained production, whatever the skill; and
3. establish the presence of creativity at levels above 0+/Novice High.

These three characteristics are the hallmarks of AEI proficiency in the applications and procedures discussed here.

The AEI proficiency framework has been applied to many evaluation situations, including assessment at the end of training or a course; for placement; before or after a stay in a target-language country; and for awarding course credit (Fischer, 1984). Although the AEI framework is often used for diagnostic purposes, its principal application has been to assess proficiency.

Testing approaches vary by skill modality and, in

some instances, by agency or school. A general outline is presented here.

All Skill Modalities. Testers of AEI proficiency use either the ILR Skill Level Descriptions or the ACTFL/ETS Proficiency Guidelines. In both, the concept of AEI proficiency is constant, requiring the testing of an examinee's ability to perform in a consistent and sustained manner for a suitable period, depending on the level.

Rating Procedures. All rating procedures ultimately depend on the definitions. In procedures using interviews, the government tends to use two raters, while academia uses one. In *direct* rating of speaking and writing, the sample is elicited and the examinee's performance is evaluated on the spot against the definitions. Oral proficiency interviews are audiotaped for possible verification later. Through listening and reading interviews, the receptive skills can also be rated immediately. *Indirect* rating occurs after paper-and-pencil tests have been normed. It may appear that tests rated in this way have no connection to real-life performances. But this is an illusion, because before they are approved as testing instruments, such tests are calibrated directly, through interviews, against real-life performances of examinees. Thus, behind the indirect proficiency test, by which a given skill is rated through a multiple-choice instrument, for example, lie real-life behaviors representative of the AEI definitions that are used to rate interviews.

Moreover, all proficiency tests require both a "floor," the level at which the examinee can consistently and sustainedly perform, and a "ceiling," the level at which the examinee no longer can sustain performance consistently. This differs from achievement tests, which have a floor but no ceiling.

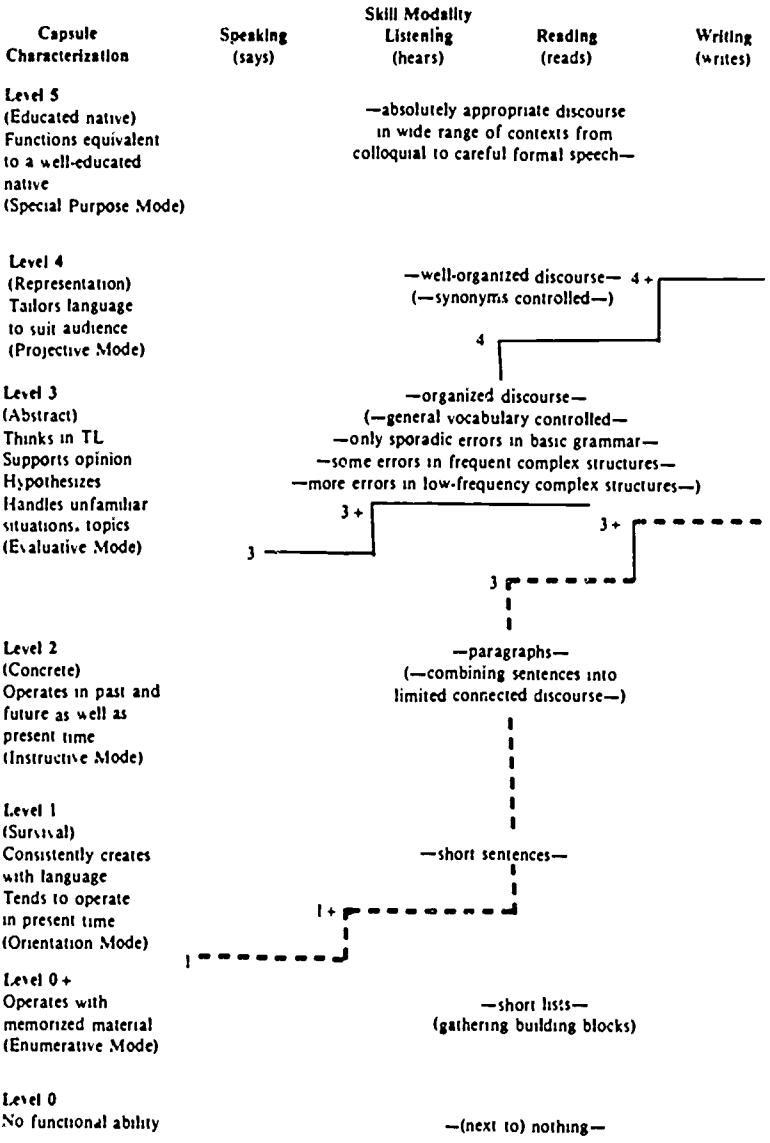
Defining Levels of the Four Skills. In a holistic, top-down view, the definitions of the four skills may be assumed; in their fullest form, they are defined as the

way in which a well-educated native speaker speaks, listens, reads, or writes. In an atomistic, bottom-up view, on the other hand, opinions vary widely on the nature of each skill. The lowest levels of reading, for example, may be represented by the ability to decode, the ability to scan for specific information, or the ability to skim for general information, among other tasks. Because atomistic definitions are difficult to interrelate with holistic ones, commonalities across skill modalities within the AEI scales are stressed instead.

In a holistic, top-down view of the four skills, there are two other reference points besides the well-educated native speaker at the apex. These reference points commend themselves naturally because they derive from a functional constellation of features. The middle reference point is the first general utility level, and the lowest reference point is the first usable level. The difference between the latter two is the degree of independence and accuracy. At the first general utility level, speakers have a level of production that is generally independent and they are confident that they will not commit egregious errors or pass misinformation. The first usable level, in contrast, represents a level of production that is reached when the examinee is no longer limited to formulaic language, and can use his or her knowledge of the language to create novel sentences.

Here the concept of profiles is useful, because the first usable level in reading falls at a different point on the scale from the first usable level in speaking (i.e., R-3 versus S-1; see Figure 1.1). ILR experience confirms the placement of these two levels for speaking, and their placement is generally confirmed for reading. ILR experience may not be extensive enough to confirm the author's experience of these two categories for listening and writing.

In sum, the AEI scales present three levels with diminishing degrees of utility referenced from the top down: the well-educated native level, the first general utility level, and the first usable level. These reference points cut across skills, but do not necessarily fall at the



Legend: Solid line = first general utility level; dotted line = first usable level.

Figure 1.1. ILR Skills Overview

Note. From " 'The' Question" by P. Lowe, Jr., 1984, *Foreign Language Annals, 17.* p. 382. Copyright 1984 by American Council on the Teaching of Foreign Languages. Reprinted by permission.

same points on the scales for every skill. Consequently, testing may take place at different levels depending on the level of proficiency desired for the skill tested and depending on the nature of the test procedure or instrument.

Testing Speaking. Testers use the oral proficiency interview, although some tasks, such as briefing and debriefing, are agency-specific. A general-language AEI interview is currently used by ACTFL, ETS, and most ILR agencies. The interview used at FSI is more exclusively job-related than at other agencies. Because a single testing procedure was used in the earliest stages of proficiency assessment, the transfer to academia was simplest for oral proficiency testing.

For the other skills, the transfer has been more difficult, with varying approaches and less widespread understanding of the techniques involved. Moreover, academia has used numerous testing approaches to listening, reading, and writing, and has defined these skills differently.

Testing Listening. From a holistic, top-down point of view, ILR users generally agree on the nature of listening, but many disagree on the contexts that a listening proficiency test should cover.[14] The lack of clear-cut guidelines for assessing these differing contexts hinders the design of listening comprehension tests. It is unclear whether the test should focus on participative listening, overheard participative listening, nonparticipative listening, or a mix of these three. The first category is distinguished from the others by the presence of opportunities to seek clarification; overheard participative and nonparticipative listening are differentiated by the nature of their content; that is, real-life, everyday situations versus lectures and broadcasts. No consensus has been reached on the issue of whether broadcast-quality tapes should be used for real-life contexts; whether natural noise should be included as part and parcel of the background; or whether a

combination should be used. Moreover, perhaps a variety of voices, genders, ages, and regional variants should be included.[15] Obviously, a lack of suitable guidelines hampers development in this area.

Some ILR agencies, like FSI, assign no listening score. Other agencies evaluate listening comprehension, but by varying methods. Some derive a listening score from the oral proficiency interview, interspersing speaking probes with listening probes. Such questions are specifically designed to test a rise in the level of listening comprehension (see Lowe, 1985b, on the listening offset). Others supplement the oral proficiency interview through a listening comprehension interview whose passages are either played on tape or read aloud to the examinee by the tester(s). The examinee verbalizes the essential meaning of the passages (in English at lower and mid levels; optionally in the target language at higher levels). At least one agency, DLI, administers a separate tape-mediated listening comprehension examination as part of its Defense Language Proficiency Tests (DLPTs). Earlier versions of the test, DLPT I and II, tested only reading and listening comprehension. The latest version, DLPT III, includes a semi-direct speaking test (see Lowe & Clifford, 1980, on the ROPE; Herzog, cited in Rose, 1987).

Testing Reading. In a holistic, top-down view, the definition of the reading skill is also straightforward. However, because current views of reading are generally atomistic and bottom-up, there is much disagreement in academia on the nature of reading proficiency. Drawing on ILR experience concerning the relative utility of the different skills at various AEI levels, the AEI top-down categories and the more atomistic bottom-up categories may be combined into the following schema, with the two sets of categories intersecting at Level 3/Superior: At the top, Level 5 is the well-educated native reader; at Level 4, the reader is at the first general utility level; and Level 3 represents the first usable level. All three categories assume automatic

comprehension, with diminishing degrees of ability in ILR functions and content, and diminishing accuracy, as the scale is descended. From the bottom up, the AEI definitions already reflect categories that are common in academia. Possible definitions range from decoding ability to automatic comprehension. The AEI definitions assume decoding ability at the 0; 0+/Novice Level, and automatic comprehension beginning with the 3/ Superior Level. The ability to puzzle out what the topic of a passage may be after multiple passes is vastly different from that of automatically comprehending the same passage in one pass at near-native reading speed. The AEI definitions assume this higher-level ability starting at Level 3/Superior, and throughout the scale they require proof of consistent and sustained ability to read (however that may be defined for the level in question). The AEI scales' assumption of such ability differs sharply from what most in academia consider as reading in a foreign language, at least at its earliest stages.

Three other factors influence bottom-up views of reading proficiency, primarily involving the fact that reading is a receptive skill and must be tested by indirect means. First, the prompt, in this case the text the examinee is supposed to understand, must be graded for AEI level(s). Unless the level of the text is determined, it is impossible to assess a level of proficiency for the examinee. While much work remains to be done in developing criteria for categorizing texts at particular levels, the ILR Testing Committee has graded more than 30 English texts to exemplify the text rating process, and it is possible to develop capsule characterizations of these text types (see Figure 1.2).[16] Doubtless other text types would lead to slightly different characterizations.

A second complication is that the rater's perception of the extent to which an examinee comprehends a reading text cannot be direct, but must occur through another channel, often a productive skill, that is, speaking or writing. If this other skill is poorly controlled, then this channel may introduce so much disturbance

that the rater may not discover the level at which the examinee truly understood the target-language text. For example, an examinee with a high level of proficiency in reading Japanese may be unable to demonstrate his or her true level of comprehension of a Japanese text when asked to paraphrase it orally in Japanese due to a low level of speaking proficiency.

A third and hotly debated factor is the contribution of real-world knowledge (Bernhardt, 1986; Lange & Lowe, 1987). Such knowledge may be presumed among well-educated native speakers. Indeed, in the government's experience, the scales are developmental. A speaker at the peak of the scale would be expected to integrate various abilities and real-world knowledge; at lower levels, less real-world knowledge is presumed. When the scale is applied to adolescents and children, however, less real-world knowledge can be expected at all levels. Hirsch (1987) goes far in answering the question of how much knowledge can be assumed with a lengthy list of cultural facts he expects educated American adults to know. Obviously, other languages place similar demands on well-educated natives (e.g., Stein's [1946] *Kulturfahrplan* for Germans).

In sum, because academia has had greater experience with reading than with speaking, opinions diverge markedly regarding the need, efficacy, and appropriateness of conducting AEI reading proficiency tests (Bernhardt, 1986).[17] Procedures for assessing reading include the reading interview used at FSI (now being revised); all-level, multiple-choice reading proficiency examinations (CIA and DLI); and level-specific tests (NSA) (Lowe, 1984a, 1984b). The reading interview usually follows the oral proficiency interview, in which case testers select reading passages at or slightly below the speaking level. If it does not follow the oral proficiency interview, the testers begin with a Level 1 passage. They then progress upward or downward, selecting the next passage as indicated by the examinee's performance and continuing iteratively until both a floor and a ceiling in the examinee's performance are

Figure 1.2. Capsule Characterizations of Written Texts

The following capsules represent the major characteristics used in grading passages according to the ILR reading proficiency scales. The following assumptions were made.

1. The descriptions would primarily treat expository prose.

2. The descriptions would outline, but not exhaustively describe, the target area into which most passages at each level fall.

3. The descriptions would be cumulative, each higher-level description subsuming those below and modifying them as necessary.

4. Straightforward texts exist whose vocabulary, structure, and organization clearly demonstrate the level in question, prove readily gradable, and consequently need not be described separately.

5. Most texts focus on a single level, but also evince variety. For example, a text graded Level 3 might nonetheless use 3^+ or 4 vocabulary or structure, but in such a way as not to prove essential to understanding the core of the passage.

6. The *major plus rule* applies to grading texts: To be considered a plus level, a passage must evince numerous features of the next higher base level but fail to use them as consistently and sustainedly as a passage from that level would. Plus levels are treated separately only at Levels 0^+ and 1^+, where attention must be drawn to their salient features.

7. The descriptions should build on the work of several ILR agencies, in this case Child (1987) and Ray T. Clifford's additions to Child's work.

8. Experience is the best teacher. Consequently, the best introduction to grading texts is a series of graded passages and commentary with sufficient variation at each level to impart a sense for the system.

CAPSULE CHARACTERIZATIONS OF WRITTEN TEXTS BY LEVEL

Clifford's *Formulaic* Mode (commonly fixed phrases and isolated words): To the extent that 0^+ texts exist, they tend to be loosely connected groups of words, such as those referring to the weather: sun, rain, and so on. Texts are often strongly supported by context, usually visual.

Child's *Orientation* Mode (main ideas): Level 1 texts contain short, discrete, simple (occasionally compound) sentences, whose vocabulary and structure are simple and whose ordering of information

may prove quite loose. Reorderings both within sentences and within paragraphs are often possible. Moreover information is not tightly packed and consequently words may be deleted without drastically altering meaning. Passages at this level may be supported by context.

Level 1+: Such texts present somewhat longer sentences; vocabulary often contains higher-level items; and sentence structure may be *compound*. Passages may be written in the past, and are usually packed with information. Material cannot be easily deleted without altering the sentence's meaning, and reordering often proves impossible as material is presented in a logical sequence compared with Level 1.

Child's *Instructive* Mode (main ideas plus supporting facts, but not details): Level 2 texts contain complex sentences, verbal times other than the present, and, in news articles, densely packed information. Vocabulary grows more topic-specific, and organization begins to reflect target-language types. Often less obvious, but more pervasive, is *shaping*—the subtle choosing of material, of ordering, and of interpretative comment on the author's part. Author assumes some shared target language culture as background.

Child's *Evaluative* Mode (main ideas, supporting facts, details, and inferences): Level 3 texts often treat abstract topics, and the language itself reflects this through abstract formulations. Texts may present author-intended inference, hypothesis, and suasion with the author often intending that the reader evaluate the material. At this level information-packing may lead to stylistics, inference and occasionally emotional aspects of the author's message. Shaping is evident as are target-language orderings that may deviate markedly from those of American English expository prose. The author assumes that readers share much target culture as background.

Child's *Projective* Mode (marked by unpredictable turns of thought): Level 4 texts demonstrate the author's virtuosity with language, often mixing registers (formal and informal, etc.), achieving subtlety and nuance, frequently evincing tone (irony, humor, etc.), cogently persuading and generally challenging the reader to follow unpredictable turns of thought. Language achieves a high level of abstraction, and author-shaping and organization is clearly evident. Author assumes reader shares target-language culture at a high level. Choice of vocabulary, idioms, and structure proves so highly appropriate that attempts at substitution produce nuances not intended by the author.

Clifford's *Special Purpose* Mode (the highest levels of language, organization, esthetics, and thought): Level 5 texts are often written for special purposes and therefore are often idiosyncratic (e.g., avant garde literature, high-level legal documents). Choice of vocabulary, structure, style, register, organization, cultural references, and so on, proves absolutely appropriate.

established. Thus, like the oral proficiency interview, but unlike more common reading tests, the reading interview is adaptive; the difficulty of the items is tailored to the examinee.

The examinee demonstrates the extent of understanding by paraphrasing the content of the passage. At the Low and Mid levels, the examinee uses English; at High levels, he or she has the option of paraphrasing in the target language.

At each level, the reading definitions state the nature of the tasks and degree of understanding required. At Level 1/Intermediate, the examinee should understand the main ideas. At Level 2/Advanced, he or she should understand the main ideas plus some supporting facts. At Level 3/Superior, the examinee should understand the main ideas and supporting facts; draw out inferences intended by the author; and perceive analogies. The level of the passages is raised until a linguistic ceiling is reached.

Although the reading interview is now an uncommon approach, it is important to note that many statements in the definitions refer to the behaviors observed during reading interviews. Some behaviors parallel those in the oral proficiency interview. For example, the definitions state that a Level 0/Novice examinee tends to understand isolated words. In an actual reading interview, a Level 0/Novice examinee typically points to individual words in a Level-1 text, the ones he or she understands in a text that is otherwise incomprehensible to him or her.

Testing Writing. Writing is the skill least tested by ILR agencies. Only two agencies, CIA and DLI, test writing, and then only for special cases (Herzog and Katz, this volume). The usual approach has been to choose a topic according to the level of proficiency required for a particular job. The resultant performance is judged against the definitions. An analogous scoring procedure used elsewhere is holistic rating (Herzog and Katz, this volume.) Many CIA jobs, for

example, require Level 3/Superior writing skills or better.

Because any writing performance is fixed in black and white, Ericson (personal communication) recommends a three-stage approach to rating:

1. Read the composition for communicative effectiveness;
2. Reread for specific factors contributing to both success and failure to communicate; and
3. Combine the results of the two approaches to derive the global writing score.

Herzog and Katz (this volume) address the difficulties of this approach.

Conclusion

This chapter has advanced probable reasons why proficiency assessment's earlier history has often gone unassimilated. To a large extent, this chapter has presented proficiency assessment in its own terms, that is, as a consistent framework, evolving over time and applying to all four skill modalities. It is hoped that this chapter provides an overarching context for the significant issues, explored in the following chapters, that have arisen as proficiency assessment has been introduced into academia.

NOTES

1. A general familiarity with the ACTFL/ETS/ILR definitions is assumed. Concise introductions are found in Lowe and Liskin-Gasparro (1986) and Lowe (1983) for speaking, and in Lowe (1984a) for other skill modalities.

2. A distinction should be drawn between a *procedure* and an *instrument*. While an instrument is invariable for all administrations, a procedure permits the administrator to vary the testing approach. Both aid assessment, but in different ways. In any

assessment procedure, the content may vary, but the performance criteria (levels, functions, and accuracy) required are constant, so that each administration at a given level furnishes a parallel test. Interviews, whether for speaking, listening, or reading, are classified as procedures. Paper-and-pencil tests of any kind are instruments.

3. The term *functions* is used here to mean fixed *task universals* that characterize specific levels in the ILR system. It is not used in the functional/notional sense of a set of variable qualifiers affecting language communication (Munby, 1978).

4. Trisections are systematic condensations of the AEI definitions, and focus on ILR functions, context/content, and accuracy for each level. (For a discussion and a chart showing the trisection for each skill, see Bragger, 1985, p. 47 [speaking]; Omaggio, 1986, pp. 129-132 [listening]; Magnan, 1985, p. 111 [writing]; Omaggio, 1986, pp. 153-156 [reading].)

5. For a more recent study, see Clark, 1987b.

6. Some differences in wording represent attempts to use the terminology more current in the literature on the nonoral skills (see section entitled "1980s Tasks in 1950s Language" later in this chapter).

7. Revision of the definitions is a continuing process both for ACTFL and the ILR, with new insights and experience being incorporated as they become available. In this sense, all sets of the definitions are provisional.

8. A clearly important third group, as yet unaddressed by the AEI definitions, is language losers—those who are in the process of forgetting or have already mostly forgotten the target language(s) (Lambert & Freed, 1982). Research is needed to determine what happens when people who know a language fail to use it; which skill modalities are most stable, and which are least stable (Lowe, cited in Lambert & Freed, 1982). An overarching question is how such performances challenge the system.

The obvious differences among these three groups—users, learners, and losers—provide a rich field for investigation. Another set of differences lies in the often disparate goals between the many government programs that are intensive and focused on teaching functional skills and some programs in academia whose articulation between program content and academia's overall goals for the learner are less clear.

9. The ILR Skill Level Descriptions' plus levels are inconsistently designated in the ACTFL/ETS versions, being referred to as plus at the Advanced level (Advanced Plus = 2^+) but as High at the lower levels (Intermediate High, Novice High = 1^+, 0^+ respectively).

10. The factors enumerated here are used by most ILR agencies, though FSI uses a slightly different set.

11. For example, in reading, instead of pronunciation, orthography; instead of fluency, speed. See Herzog and Katz, this volume, for factors contributing to writing, such as organization and discourse methods.

12. This discussion complements Lowe, 1985b and 1986.

13. For a possibly more atomistic, bottom-up view of proficiency and applicable testing procedures, see Larson and Jones, 1984.

14. For details on listening, see Douglas (1987); Foreign Language Annals (September 1984), Joiner (1986). In addition, many of the statements made here about reading may well apply to listening, which is related as a receptive skill.

15. For additional clarificiation of these problems, see Valdman (1987).

16. The accompanying figure builds on the work of Child, Clifford, Herzog, and Lowe, and outlines primarily the characteristics of exemplary expository prose tests at each ILR level.

17. Several AEI reading proficiency tests have been developed in academia: Japanese Proficiency Test and Russian Proficiency Test (ETS); Chinese Proficiency Test (Center for Applied Linguistics); and the Hindi Proficiency Test (University of Pennsylvania).

References

American Council on the Teaching of Foreign Languages. (1986). *ACTFL proficiency guidelines*. Hastings-on-Hudson, NY: Author.

American Council on the Teaching of Foreign Languages. (1982). *ACTFL provisional proficiency guidelines*. Hastings-on-Hudson, NY: Author.

Bachmann, L.F., & Clark, J.L.D. (1987). The measurement of foreign/second language proficiency. *Annals of the American Political and Social Sciences, 490* (March), 20-33.

Bachmann, L.F., & Savignon, S.J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFl oral interview. *Modern Language Journal, 70*, 291-297.

Bernhardt, E. (1986). Proficient texts or proficient readers? *ADFL Bulletin, 18*(1), 25-28.

Bragger, J.D. (1985). The development of oral proficiency. In A.C. Omaggio (Ed.), *Proficiency, curriculum, articulation: The ties that bind.* Middlebury, VT: Northeast Conference.

Buck, K.A. (1984). Nurturing the hothouse special: When proficiency becomes performance. *Die Unterrichtspraxis, 17,* 307-311.

Byrnes, H. (1987a). Discourse structure and communicative strategies in the development of communicative ability. In A. Valdman (Ed.), *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency.* Bloomington, IN: Indiana University.

Byrnes, H. (1987b). Second language acquisition: Insights from a proficiency orientation. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations, and concepts.* Lincolnwood, IL: National Textbook Co.

Child, J. (1987). Language proficiency and the typology of texts. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementation, and concepts.* Lincolnwood, IL: National Textbook Co.

Clark, J.L.D. (1978). *Direct testing of speaking proficiency: Theory and application* [Proceedings of a joint ETS/ILR/GURT conference]. Princeton, NJ: Educational Testing Service.

Clark, J.L.D. (1987a). Sociolinguistic aspects of communicative ability: The issue of accuracy vs. fluency. In A. Valdman, (Ed.). *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency.* Bloomington, IN: Indiana University.

Clark, J.L.D. (1987b). A study of the comparability of speaking proficiency interview ratings across three government language training agencies. In K.M. Bailey, T.L. Dale, & R.T. Clifford (Eds.), *Language testing research.* Monterey, CA: Defense Language Institute.

Clifford, R.T. (1980). FSI factor scores and global rating. In J.R. Frith (Ed.), *Measuring spoken language proficiency*. Washington, DC: Georgetown University Press.

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), *Language proficiency and academic achievement.* Clevedon/Avon, England: Multilingual Matters.

Douglas, D. (1987). Testing listening comprehension. In A. Valdman, (Ed.). *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency.* Bloomington, IN: Indiana University.

Educational Testing Service. (1981). *A common metric for language proficiency* (Final Report for Department of Education Grant No. G008001739). Princeton, NJ: Author.

Educational Testing Service. (1982). *ETS oral proficiency testing manual.* Princeton, NJ: Author.

Fischer, W.B. (1984). Not just lip service: Systematic oral testing in a first-year college German program. *Die Unterrichtspraxis, 17,* 225-39.

Frith, J. (1979). Testing the testing kit. *ADFL Bulletin, 11* (2), 12-14.

Garrett, N. (1986). The problem with grammar: What kind can the language learner use? *Modern Language Journal, 70,* 133-48.

Higgs, T.V. (1985). Teaching grammar for proficiency. *Foreign Language Annals, 18,* 289-96.

Higgs, T.V., & Clifford, R. T. (1982). The push towards communication. In T.V. Higgs (Ed.), *Curriculum, competence and the foreign language teacher.* Lincolnwood, IL: National Textbook Co.

Hiple, D.V. (1984). The ACTFL proficiency projects: An update. *Die Unterrichtspraxis, 17,* 327-29.

Hiple, D.V. (1986). A progress report on ACTFL proficiency guidelines, 1982-1986. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations, and concepts.* Lincolnwood, IL: National Textbook Co.

Hiple, D.V. (1987). *The dynamics of organizational growth and development in an association providing adult continuing professional education.* Unpublished doctoral dissertation, State University of New Jersey, Rutgers, NJ.

Hirsch, E.D., Jr. (1987). *Cultural literacy: What every American needs to know.* Boston: Houghton Mifflin.

Interagency Language Roundtable. *ILR skill level descriptions.* (1985). Washington, DC: Author.

Joiner, E.G. (1986). Listening in the foreign language. In B.H. Wing (Ed.), *Listening, reading, writing: Analysis and application.* Middlebury, VT: Northeast Conference.

Jones, R.L. (1975). Testing language proficiency in the U.S. government. In R.L. Jones & B. Spolsky (Eds.), *Testing language proficiency.* Arlington, VA: Center for Applied Linguistics.

Jones, R.L., & Spolsky, B. (Eds.). (1975). *Testing language proficiency.* Arlington, VA: Center for Applied Linguistics.

Lambert, R.D., & Freed, B.F. (Eds.). (1982). *The loss of language skills.* Rowley, MA: Newbury House.

Lange, D., & Lowe, P., Jr. (1987). Grading reading passages according to the ACTFL/ETS/ILR proficiency standard: Can it be learned? In K. Bailey & R.T. Clifford (Eds.), *Proceedings of the pre-TESOL testing colloquium.* Monterey, CA: Defense Language Institute.

Larson, J.W., & Jones, R.L. (1984). Proficiency testing for the other language modalities. In T.V. Higgs, (Ed.), *Teaching for proficiency: The organizing principle.* Lincolnwood, IL: National Textbook Co.

Liskin-Gasparro, J.E. (1984a). The ACTFL proficiency guidelines: A historical perspective. In T.V. Higgs (Ed.), *Teaching for proficiency: The organizing principle.* Lincolnwood, IL: National Textbook Co.

Liskin-Gasparro, J.E. (1984b). The ACTFL proficiency guidelines: Gateway to testing and curriculum. *Foreign Language Annals, 17* , 475-89.

Lowe, P., Jr. (1980, March). *Gestalts, thresholds, and the ILR definitions.* Paper given at the 1980 presession sponsored by Georgetown University Round Table on Languages and Linguistics and the Interagency Language Roundtable, Washington, DC.

Lowe, P., Jr. (1983). The ILR oral interview: Origins, applications, pitfalls, and implications. *Die Unterrichtspraxis 16*(2), 230-244.

Lowe, P., Jr. (1984a). Setting the stage: Constraints on ILR receptive skills testing. *Foreign Language Annals, 17* , 375-79.

Lowe, P., Jr. (1984b). "The" question. *Foreign Language Annals, 17,* 381-88.

Lowe, P., Jr. (1985a). *The ILR handbook on oral interview testing.* Washington, DC: DLI-ILR Oral Interview Project.

Lowe, P., Jr. (1985b). The ILR proficiency scale as a synthesizing research principle: The view from the mountain. In C.J. James, (Ed.), *Foreign language proficiency in the classroom and beyond.* Lincolnwood, IL: National Textbook Co.

Lowe, P., Jr. (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz and, particularly, to Bachman and Savignon. *The Modern Language Journal, 70,* 391-97.

Lowe, P., Jr. (1987). Four challenges to proficiency: Comments from within the AEI proficiency framework. In A. Valdman (Ed.), *Proceedings of the Symposium on Evaluating Foreign Language Proficiency.* Bloomington, IN: Indiana University.

Lowe, P., Jr., & Clifford, R.T. (1980). Developing an indirect measure of overall oral proficiency. In J.R.

Frith (Ed.), *Measuring spoken language proficiency.* Washington, DC: Georgetown University Press.

Lowe, P., Jr., & Liskin-Gasparro, J.E. (1986).*Testing speaking proficiency: The oral interview* [ERIC Q&A]. Washington, DC: Center for Applied Linguistics.

Magnan, S.S. (1985). Teaching and testing proficiency in writing: Skills to transcend the second language classroom. In A.C. Omaggio (Ed.), *Proficiency, curriculum, articulation: The ties that bind.* Middlebury, VT: Northeast Conference.

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

Murphy, C., & Jiménez, R., (1984). Proficiency projects in action. In T.V. Higgs, (Ed.), *Teaching for proficiency, the organizing principle.* Lincolnwood, IL: National Textbook Co.

Omaggio, A.C. (Ed.). (1985). *Proficiency, curriculum, articulation: The ties that bind.* Middlebury, VT: Northeast Conference.

Omaggio, A.C. (1986). *Teaching language in context: Proficiency-oriented instruction.* Boston: Heinle & Heinle.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). *The measurement of meaning.* Urbana, IL: University of Illinois Press.

Rice, F.A. (1959). Language proficiency testing at the Foreign Service Institute. *The Linguistic Reporter, 1,* 4.

Rose, M.G. (Ed.), (1987). *Translation excellence: Assessment, achievement, maintenance* (American Translators Association Scholarly Monograph series, Vol. 1). Binghamton, NY: University Center at Binghamton (SUNY).

Schulz, R.A. (1986). From achievement to proficiency through classroom instruction: Some caveats. *Modern Language Journal, 70,* 373-79.

Shohamy, E. (1987). Comments. In A. Valdman (Ed.), *Proceedings of the Symposium on Evaluating Foreign Language Proficiency.* Bloomington, IN: Indiana University.

Sollenberger, H.E. (1978). Development and current use of the FSI oral interview test. In J.L.D. Clark (Ed.), *Direct testing of speaking proficiency.* Princeton, NJ: Educational Testing Service.

Stein, W. (1946). *Der Kulturfahrplan.* Berlin-Grunewald: F.A. Herbig.

Valdman, A. (1987). The problem of the target model in proficiency-oriented language instruction. In A. Valdman (Ed.), *Proceedings of the Symposium on Evaluating Foreign Language Proficiency.* Bloomington, IN: Indiana University.

Wilds, C.P. (1975). The oral interview test. In R.L. Jones & B. Spolsky (Eds.), *Testing language proficiency.* Arlington, VA: Center for Applied Linguistics.

Wilds, C.P., & Jones, R. (1974). *Interagency interrater interreliability study* (Internal document). Washington, DC: Foreign Service Institute; Central Intelligence Agency.

II
A Research Agenda

by John L.D. Clark and John Lett,
Defense Language Institute

So much has been said about proficiency-based testing within the past few years that the reader may wonder what else can possibly be learned or reported about it. Fortunately for the researcher in this area—and also ultimately for the user of the information provided by this research—it is unlikely that a point will ever be reached at which "enough" systematic and meaningful data will have been gathered about the strengths and weaknesses, appropriate and inappropriate applications, valid and less valid interpretations, and other major characteristics of proficiency-based testing approaches. While it is not possible to provide a comprehensive inventory of potential research activities that might be carried out over the next several years, a number of issues can be brought to the reader's attention that seem to warrant high priority on the proficiency testing research agenda. Although the present focus is on oral proficiency testing, the research concepts and challenges addressed can be considered substantially applicable to assessment in the other three skill areas—listening, reading, and writing—as well.

The oral proficiency interview is examined first from the two fundamental perspectives of validity and reliability, each of which is seen to subsume and suggest a number of major empirically based studies. A discussion follows of scaling and measurement

considerations at issue in the interview procedure and associated rating process. Finally, a recently proposed comprehensive program of proficiency testing research is described that is based on an expanded model of communicative language ability that could serve as an overall organizing structure and vehicle for conducting virtually all the research activities discussed here.

Validity Considerations in Proficiency Testing

The crucial question in assessing the validity of any type of test is simple and straightforward: "How well does the test do what it is supposed to do?" With regard to tests of general proficiency in a second language, it can reasonably be assumed that such tests are supposed to determine "how well a person can speak (or understand speech in, read, or write) the language." While the meaning of "speak" (or understand, read, write) in this context is fairly clear, the proper interpretation and application of the terms "how well" and "the language" are considerably more complex, and go to the heart of the validity issue.

To address first what is meant by "the language" in a proficiency testing context, it is usually understood and agreed that "proficiency" measurement involves determining the examinee's ability to perform in a linguistically (and sociolinguistically) appropriate manner within a variety of language-use situations encountered in real-world contexts external to the instructional setting per se. Although this orientation does not rule out the possibility that a given instructional program would itself emphasize this type of language practice and development on the student's part, it does require that the test developer look exclusively at language use in "real-world" settings in specifying both the linguistic

content and manner of operation of the test.

Obviously, no single proficiency test of any reasonable length could more than merely sample the virtually unlimited range of language and language-use situations potentially encountered in the "real world." The science and art of proficiency testing thus involves, to a great extent, the judicious and informed selection of language content and testing techniques in such a way as to maximize the extent to which one can legitimately infer from an examinee's test performance a similar manner and quality of performance in a variety of real-world settings that the test is designed to reflect.

One approach is to restrict the range of real-world situations to which a given test is intended to apply and to focus test content and testing procedures on the smaller area. This strategy was adopted by Munby (1978), who developed a procedure for specifying the linguistic and sociolinguistic situations in which a particular examinee—for example, a waiter in a summer resort—would need to function. Although a language performance test focusing on such a delimited area could readily be developed, its use and interpretation would by the same token necessarily be restricted to the given area. In view of both the economic and administrative infeasibility of developing and using separate performance measures for each of the virtually unlimited number of real-world situations that could be proposed as areas for direct, specialized testing, the only viable alterative would seem to be to use a considerably smaller number of more general tests designed in such as way as to permit reasonable assumptions about the level and quality of examinee performance across a wider range of language-use contexts. This approach has the potential drawback of reduced measurement accuracy by comparison with specifically tailored performance tests.

Given the wide-scale use of the ACTFL/ETS/ILR (abbreviated here to AEI) speaking proficiency interview and associated scoring scale in both academic and government settings, it would be useful to consider,

from both conceptual and empirical research standpoints, the extent to which the interview process adequately samples the communicative situations and associated linguistic contexts in which nonnative speakers are most frequently required to perform. In this regard, an immediate observation is that the conversational portion of the AEI interview, while constituting a highly realistic sample of polite, reasonably formal conversation between relative strangers, does not lend itself well to exemplifying discourse situations in which, for example, the examinee is in a higher-status role than his or her interlocutor; is engaging in relaxed conversation with friends; or is communicating in a highly emotional or other affect-laden manner. Although an approximation of these other types of discourse settings is offered by the use of printed cards describing various communication situations that the examinee is asked to role-play with the interviewer, this approximation remains at a considerable physical and psychological remove from the genuine communicative situation.

A second concern is that neither the regular interview portion nor the situational role-play provides a ready or straightforward opportunity for the examinee to engage in extended narration, although experienced interviewers are often able to elicit some "monologue" discourse from the examinee. Also, the testing procedures used by the Foreign Service Institute (FSI) have recently been modified to include a "briefing" section in which the examinee is required to speak continuously for several minutes about information provided in a printed text studied a few minutes previously. It remains fair to say, however, that although the role-play procedure, briefing technique, and other modifications of the basic interview represent useful attempts to go beyond the sociolinguistic and discourse constraints of "polite conversation," these have been fairly adventitious and have not been guided by a preliminary analysis of the various situations (and their relative frequencies) in which nonnative learners of the language

would characteristically be required to communicate.

One way to address this aspect of the validity question would be to conduct a Munby-type (1978) analysis, at a suitable level of generality, of the communication situations most frequently encountered by given groups of nonnative users of the language. Such an analysis would produce information on which to base the content of role plays and possibly other components of the interview, depending on the group. Americans traveling abroad as tourists might constitute one such group, while another group might consist of native English speakers studying or working in the target-language country over extended periods. Analyses of the speaking requirements of these and several other groups could be conducted, and the resulting language-use profiles could be compared. To the extent that the profiles overlap, the use of a single testing process and content to apply to several different groups would be a reasonable approach. On the other hand, it would not be appropriate to attempt to apply a single uniform process and content across groups whose language-use needs are found to differ appreciably as to types of language and the situations in which the language would need to be used.

In other words, such linguistic needs analyses, applied as content/process criteria in the review of a given proficiency test, would serve to indicate and document the range of applicability of the instrument in question. At the same time, such analyses could identify real-world contexts for which the test content and manner of operation would not represent a reasonable sample; in such instances, examinee performance on the test would not provide an accurate prediction of real-life performance.

Turning to issues involved in determining "how well" the examinee is able to use the language in real-life communication settings, it is proposed for purposes of initial discussion that development and validation of the AEI—or any other scale purporting to define and quantify second language proficiency—

should take adequately into account *native-speaker judgments* of the nature and effectiveness of the examinee's communication attempts, because it is precisely with native interlocutors that the examinee would be expected to communicate in the real-life contexts the scale is intended to reflect. It should be noted that scale development based on native-speaker judgments would need to take into consideration a number of sources of variance inherent in such judgments, including the sociolinguistic context in which they are made; the age, sex, educational level, and profession of the native-speaker judge; and whether the impact of these factors varies according to the linguistic and cultural norms of the language and national community to which the judge belongs. These and related issues are discussed later in this section.

Native-speaker judgment studies available to date suggest that this is a fertile field of inquiry with major implications for proficiency-based measurement. From a strict comprehension-of-message point of view, Burt (1975) found that errors by nonnative English speakers involving incorrect word order or other sentence-level distortions resulted in considerably greater miscommunication than did "local" errors such as omission of the auxiliary in phrases such as "Why [do] we like . . . ?" Similarly, Chas in (1980) found, though in a writing rather than a speaking context, that native speakers of Spanish were much more tolerant of simple "grammatical" errors (e.g., use of the wrong gender of definite and indefinite articles) than they were of errors that affect meaning, such as the use of a contextually inappropriate lexical item. With regard to affective reactions of native speakers to various types of nonnative speech, Raisler (1976) found that spoken English that was proficient in all respects except for accent was not only negatively received by native listeners but also considered, quite erroneously, to contain "grammar errors." In a study by Galloway (1980), native speakers of Spanish viewing videotapes of American learners attempting to communicate in that language gave

considerable credit to people whom they perceived as exerting "effort in expressing themselves," by comparison with other more linguistically advanced students for whom such effort was less apparent. Galloway concluded that "an individual's lack of grammatical accuracy may not produce negative reactions if the desire and urgency to communicate are evident" (p. 430). (See also Mattran, 1977; Ludwig, 1982).

Intriguing questions can be posed about the nature and degree of "fit" between the AEI level ratings of given examinees and native-speaker perceptions of their communicative effectiveness. An immediate area of investigation would be the possibility of rating "disconnects" between various portions of the AEI scale and native speakers' quantified judgments. Would native speakers indeed make distinctions between, for example, examinees rated Novice-Mid and those rated Novice-High, or would they assign both to a single undifferentiated group? In the latter case, some concern might emerge as to whether the "fine-grained" ACTFL levels are indicative of true differences in communicative performance in the perception of native speakers.

Assuming research designs that provide adequate control of native-speaker variables, detailed linguistic analysis of "less proficient" and "more proficient" nonnative speech as judged by native listeners could be carried out both for its own sake and as one approach to validating the AEI level descriptions in a noncircular manner. For example, a set of interview samples might be judged by naive native speakers to constitute increasingly "proficient" performances; if it were also found to exhibit the types and degrees of control of structure, lexicon, phonology, and so on, that would be predicted for them on the basis of the AEI descriptions, considerable confidence could be placed in the operational value of the level descriptions as reflective of real-life linguistic phenomena. Identification of linguistic aspects that do not pattern similarly on the two continua might suggest desirable changes in wording for one or more of the level descriptions, including possibly some

shifting of the points on the AEI scale at which given linguistic phenomena are considered to occur.

While close examination of the AEI scale through the optic of naive native-speaker judgments is a research area of extremely high potential, a properly conducted research program in this area would need to address several conceptual and operational complexities. First is the degree of consistency and uniformity that native speakers would demonstrate in their judgments of language-related phenomena. If a given native speaker's appraisal of the level of proficiency represented by a particular speech sample were found to vary widely from one judging occasion to the next (i.e., low intrarater reliability), or if several native speakers, working independently, were found to assign widely differing ratings to the same speech samples (i.e., low interrater reliability), meaningful comparisons between "native judgment" data and AEI level assignments would be extremely difficult or impossible to make.

Second, closely related to the reliability-of-judging issue is the matter of providing valid and meaningful judging instructions. In this regard, references to "grammar," "vocabulary," "fluency," and other linguistic features would probably not be easily interpreted by the naive native speaker and give rise to highly idiosyncratic judging performances. Instructions to assign given speech samples to one of several arbitrary categories of "increasing proficiency" (the so-called Q-sort technique) would be expected to provide more meaningful and reliable data, as would the forced choice between pairs of speech samples (this process, however, would be considerably more laborious and time-consuming).

A third very significant issue is the distinction between linguistic and nonlinguistic characteristics of nonnative speakers' communicative performances as these might influence the judgments of native speakers. Variations in the psychological or personality characteristics of given individuals (e.g., taciturn vs. garrulous, outgoing vs. withdrawn) may be expected to

influence native speakers' estimates of linguistic proficiency to some extent, independently of the technical accuracy of the language performance. Ethnic differences between nonnative speakers and native judges, as well as real or perceived differences in social status, degree of education, and so on, as these traits are conveyed aurally and/or visually in the course of the judging process, would be expected to introduce a certain degree of error variability in the rating of linguistic proficiency per se.

A fourth concern is that examinees vary widely in the extent to which they are perceived as effective communicators in their native language. At least some of this variance would presumably be reflected in appraisals of their target-language competence by native judges. Although it would be appropriate for this variance to contribute to the results of tests aimed at, for example, selecting the individual best suited for a job in which interpersonal communication skills were a primary consideration, it would seem desirable to limit such variance when the objective of the judgment is simply to assess the extent to which an individual has acquired the ability to use a second linguistic system. One approach lending itself to exploratory research would be to administer the AEI interview to examinees in both the target language and the native language, statistically partial out the effects of the latter on the former, and examine the resulting scores from a variety of perspectives.

Despite the complexities in the use of native-speaker judgments as external criteria for developing language proficiency tests, and the associated need to pay close attention to these factors in the design and conduct of research studies, none of these issues should be considered to pose insurmountable problems or negate in any way the extremely important role that "native-speaker studies" would be expected to play in further investigations of the validity and practical utility of the AEI proficiency scale and rating technique.

The discussion so far has dealt primarily with the

substance and manner of operation of the AEI testing procedure itself, or with questions of *content* validity. Another large potential research area is that of the *construct* validity of the AEI testing approach. Briefly characterized, construct validity studies involve the empirical determination of the tenability of hypotheses derived from measurement claims made by or implicit in the testing procedures under study. With regard to the AEI interview and scoring scale, probably the most salient hypothesis, arising directly from the stated purpose of this testing procedure, is that examinees who score at a given level on the AEI scale will in fact be able, in the real-world setting, to carry out each of the language-use tasks or functions at issue in that proficiency level description.

One theoretically possible way to test this hypothesis is constant surreptitious observation of examinees at given score levels going about their day-to-day language-use activities in the target language. This approach, unfortunately, is impracticable, despite its conceptual appeal. A second considerably more feasible approach is to have the examinee provide a self-report of ability to perform the indicated tasks or functions within the actual language-use setting. Self-report questionnaires on perceived degree of functional ability in the language have been developed and used with some success in a parametric study of second language learning by Peace Corps volunteers (Carroll, Clark, Goddu, Edwards, & Handrick, 1966) and in projects of the Experiment in International Living (1976). More recently, an Educational Testing Service study found that a set of "can-do" self-appraisal statements correlated at a level of about .60-.65 with results of an AEI-type direct proficiency interview (Clark, 1981).

Another criterion data-gathering possibility is to ask second parties who are in a reasonable position to do so to make proficiency-related judgments about examinees' language performance in specified contexts. For example, supervisors of foreign service officers might be asked to make judgments of these individuals' ability

to perform the particular communicative tasks at issue in the upper range of the AEI scale, with these judgments being subsequently correlated with the officially measured proficiency levels.

In all of these construct validation efforts, it would of course be necessary to keep in mind the potential unreliability of the self-rating, second-party, or other types of criterion data themselves, and to refine and objectify as much as possible the questionnaires or other elicitation procedures used in obtaining the validation information. While "perfect" criterion data will probably never be obtained, this consideration should in no way vitiate the concept of construct validation as applied to language proficiency testing, nor diminish the practical need for such validation. To the contrary, the dearth of construct validation studies within AEI research to date suggests a high priority for such investigations.

Reliability Considerations

Issues of *reliability* also figure prominently in the program of research needed to more fully investigate and develop both the AEI technique and other approaches to language proficiency assessment. Just as there are several different aspects of validity, each contributing to an appreciation of the overall "validation situation" for a particular instrument or testing approach, several different types of "reliability" exist, each meriting examination with a comprehensive analysis of this aspect of test performance. To begin with a generalized definition of the term, a test can be viewed as reliable to the extent that it provides identical scoring results for a given individual across each of a number of varying test-administration or other conditions (and under the assumption that the individual's true knowledge or ability in the subject matter remains constant across the testing occasions in question). For the AEI interview, major "varying conditions" would include:

1. variation in the types of discourse, topics, and so on broached in the interview producing differences in performance attributable solely or predominantly to the "luck of the content draw";
2. variation in interviewing technique, including differences in the personality or "style" of the interviewers, that affects performance independently of overall proficiency; and
3. for any given interview performance, variation in ratings assigned by the judge(s), either for a given judge rerating a particular interview (intrarater reliability), or across two or more judges rating the same performance (interrater reliability).

The potential for scoring unreliability associated with the first item might be reduced in at least two ways. First, the length of the interview could be extended so that a wider range of discourse styles and topical areas can be explored. However, given the already considerable human-resource requirements of interview-based testing of "normal" duration (about 10-40 minutes, depending on the overall proficiency level of the examinee), lengthening of the interview would probably not be well received from an administrative or budgetary perspective. A more viable approach would be to standardize the content of the interview more closely, both across examinees and interviewers, to reduce or eliminate the content sampling problem as a potential source of unreliability. At one extreme of the degree-of-standardization continuum is the completely fixed format of the so-called "semidirect" speaking test (Clark, 1979), in which the examinee orally responds to questions posed by a tape recording and/or to stimulus material in a printed booklet—an approach that completely eliminates across-test variation in test content. Semidirect tests, designed to reflect as closely as possible the AEI interview with respect to both the topical areas dealt with and types of discourse elicited, have been developed (Clark, 1986) and have been found to correlate significantly (r = .89-.96)

with concurrent live interviews. However, because the semidirect approach does not provide for the examinee to negotiate meaning, use repair strategies, or carry out other interactive discourse tasks characteristic of live conversation, the face and operational validity of this testing procedure is somewhat less than that of the live interview.

A combined approach is possible that would both maintain the interactive nature of the live interview and increase the uniformity of test content: The tester could use printed scripts or other aids as a guide in selecting particular types of questions or topics to be broached in the interview. This general approach has already been implemented to some extent in the "situation cards" developed by ACTFL, which provide descriptions of potential role-play scenarios at various AEI scale levels. Analogous lists of possible questions or categories of questions from which the tester could draw as necessary during the conversational portion of the interview could also help increase the across-interview uniformity of content and testing procedure. Cautions to heed when using this approach include both the need to provide a sufficient number of questions in the selection pool to keep details of the test content from becoming known to prospective examinees, and the need to maintain a reasonable degree of flexibility and spontaneity of discussion within the test as a whole.

Across-interview variation in examinee performance attributable to differences in testers' elicitation techniques—including affective or personality characteristics of the interviewer—may also be presumed to negatively affect test-retest reliability. Interviewers who show genuine interest in the information being communicated and who interact with the examinees in an empathetic manner would, in general, be expected to elicit better performances from the same examinees than would more disinterested or "colder" interviewers, especially at the lower range of the proficiency scale. In view of the fairly subtle phenomena at issue, an empirical test of this hypothesis would probably require

an exaggerated approach in which participating interviewers would consciously emphasize either a highly interested and friendly or highly disinterested and cool testing style in interviewing the same group of examinees. Counterbalanced administration of the two types of interview, together with single- or double-blind ratings, would be needed to control for test-order and other effects, and a fairly large number of examinees also would be required. However, there are no insurmountable research-design impediments to a thorough examination of these and other variables in interviewer and elicitation techniques as these may relate to differences in examinee performance across different interview occasions.

A major program of empirical research is needed to fully examine the intra- and interrater reliability characteristics of the AEI interview and scoring scale. Although a few small-scale studies have been conducted in this area (Adams, 1978; Clark 1978, 1986; Clark & Li, 1986; Shohamy, 1983), little definitive information is presently available in answer to questions such as the following: For trained raters, what is the typical amount of scoring variation that a given rater will exhibit in the repetitive scoring of a given examinee performance? Secondly, how and to what extent does intrarater reliability vary as a function of the technical mode (audiotape versus videotape) under which the rerating takes place? It may be hypothesized that reratings based on audiotape playback will, in general, result in somewhat lower ratings than were initially assigned. Such a phenomenon would be attributable to the fact that grammatical and other linguistic shortcomings in the speech sample would be more salient in a relaxed, after-the-fact, audio-only review than they would have been during the real-time interview, in which gestures or other visual cues on the part of the examinee might have masked (and perhaps properly so, from a communicative standpoint) certain technical flaws in the examinee's speech performance per se. In addition, in interviewing situations in which a

single tester is required to both elicit and rate the speech sample, the cognitive and performance demands of the former activity may make it difficult for the tester/rater to attend as closely as would otherwise be possible to specific linguistic details of the examinee's performance. For check-rating purposes, videotape (rather than audiotape) playback would more closely approximate the original communicative situation and would be considered preferable for this reason alone. If audiotape-based check-ratings were found to be systematically biased by comparison with the original ratings—again presumably in the direction of lower scores—this would even more strongly suggest the advisability of archiving and check-scoring interviews through means of videotapes rather than audiotapes.

A third question is, for situations in which a single individual is required to both administer the interview and at the same time listen attentively to and ultimately rate the examinee's speaking performance, what effect would this dual testing task have on rating reliability, by comparison with a two-tester procedure in which one person would concentrate on elicitation and the second on performance analysis and rating? Presumably, the cognitive and performance demands on the interviewer of eliciting a proper and complete speech sample would make it more difficult to attend to specific linguistic details of the examinee's performance than would be the case with a separate listener/rater concentrating only on these aspects. Thus, strictly from a scoring reliability standpoint, the listener/rater would probably be in a somewhat better position to make consistent rating judgments than the interviewer. On the other hand, the validity of this process could be questioned to the extent that the interviewer/rater attended more carefully to particular strengths or weaknesses in the examinee's linguistic performance than would be the case under normal communication situations in real life. In any event, the first step in examining questions of this type would be to conduct a comparative study of these two administration formats to determine the

nature, extent, and probable practical significance of any rating reliability differences.

Fourth, an important question that has yet to be systematically addressed concerns changes in scoring reliability on the part of testers following their initial interviewer/rater training. It is generally assumed, and indeed borne out in informal experience, that certified testers tend to exhibit over time at least some degree of departure in rating performance from the original standards against which they were trained. However, the direction and magnitude of these scoring variations, as a function of both the amount of elapsed time following initial training and the amount of interviewer/rating activity engaged in over that time period, have not been rigorously investigated. One possible research approach would be to have trained raters periodically rerate a set of calibrated interviews (possibly in alternate forms for the different rating occasions) covering the full range of AEI score levels. They would also be asked to keep a chronological log of the real interviews they were conducting and rating, as well as any other interviewing/rating-related activities (such as check-rating of colleagues' interviews) in which they were engaged during the study. Assuming a sufficiently large number of participants, it would be possible to place each person along a broad continuum of post-training activity, ranging from virtually constant interviewing and rating to little or no activity following initial training. Various correlative data could also be gathered, including the duration and intensity of the initial training; the degree of accuracy in rating the initial calibration tapes (ranging from borderline to consistently on target); trainer judgments as to how thoroughly the individual appeared to have understood and "internalized" the basic rating concepts; the tester's own level of proficiency in the target language, specifying whether the tester is a native or nonnative speaker; and, to the extent reasonably possible, several different measures of psychological or personality characteristics that might be hypothesized

to contribute to or militate against the maintenance of rating accuracy over time.

Such a study could provide data critical to arriving at informed answers to questions concerning:

1. changes in rating reliability over time in the absence of interviewing/rating opportunities;
2. the direction of these changes (i.e., tendency to increasing severity, generosity, or random variation); comparability of these changes across raters (as opposed to idiosyncratic or unpredictable variation);
3. possible interactions between maintenance or loss of scoring accuracy and the particular proficiency levels evaluated; for example, level 3 or higher performance might continue to be accurately rated, with wider departures at the lower levels (or vice versa);
4. the contribution of ongoing interviewing/rating activities to maintenance or loss of rating accuracy;
5. performance during tester training as a predictor of maintenance or loss of rating accuracy after training;
6. native versus nonnative speaker status as a predictor of maintenance or loss of accuracy; for nonnatives, the influence of the rater's own proficiency level in the target language; and
7. the relationship of background and/or personality characteristics of raters to maintenance or loss of accuracy.

Answers to such questions could lead in turn to the development of more effective guidelines for the selection of prospective interviewer/raters; improved initial training procedures; specifications for the optimum amount and frequency of posttraining rating practice needed to maintain accuracy; and more precisely designed and targeted "refresher/retraining" procedures.

A reliability-related question with major practical implications is that of *optimum interview length*. An earlier study in this area (Clark, 1978) showed quite high correlations between ratings based on severely

shortened interviews (about 6 minutes on average) and longer "standard" interviews of 20 minutes or more. Although the need for face and content validity—as well as user acceptance of the results—are likely to require a somewhat longer speech sample, it may be hypothesized that for each score level on the proficiency scale a rough total time limit exists beyond which scoring reliability can increase very little if at all. Should this hypothesis be confirmed and the limits identified, longer interviews could be said to waste expensive tester resources without increasing rating accuracy.

Although several research approaches to this question are possible, one reasonable procedure would be to have several raters independently listen to a set of normal-length interview tapes. At two-minute intervals, the raters would be asked to make their best current guess as to the examinee's proficiency level. In addition, they would be asked to indicate precisely when they become "certain" of the level they would ultimately assign. Listening would continue for several minutes beyond this point to provide an opportunity for any change of mind. For each rater, examination of the scores assigned at each of the checkpoints would indicate, for each proficiency level, the length of interview at which the individuals' rating judgments tended to stabilize. Across-rater comparisons at each time period would reveal changes in interrater reliability (if any) as a function of increasing test length, as well as possibly suggest an "upper bound" length (for each proficiency level) beyond which little or no further improvement in scoring reliability would be expected.

Scaling and Measurement Considerations

The preceding sections have examined the AEI testing procedure from the twin perspectives of validity

and reliability, each of which was seen to generate a number of secondary issues meriting further research. Although the issues of scaling to be addressed in this section can also be viewed as components of validity and reliability in the broadest sense of these terms, it is convenient to separate them for discussion here. Two basic scaling-related questions are addressed: first, to what extent can the AEI level descriptions be considered a scale in the rigorous psychometric meaning of this term? Secondly, what are the properties of this scale and the implications of these properties for the accurate measurement of communicative performance?

In order for a system of ratings or descriptions to be considered a scale, the ratings or descriptions must (a) denote the relative presence or absence of some substance or quality and (b) be capable of validly and reliably ordering people or objects according to the extent to which they possess the substance or quality in question. For example, an attitude scale that purports to rate individuals according to their degree of ethnocentricity must be able to demonstrate that it can (a) denote or label individuals possessing varying degrees of "ethnocentrism" and (b) validly and reliably (across varying testing instruments and raters) order individuals who are known—from some source of information external to that provided by the scale itself—to differ along the dimension of "ethnocentrism." These definitions assume that the dimension of "ethnocentrism" is indeed scalable and that the construct underlying the scale is essentially unidimensional.

Before discussing in detail a number of scale-related issues in the AEI proficiency level descriptions, it is necessary to acknowledge that over a period of about 30 years, language training professionals, chiefly within the U.S. government, have effectively used the ILR proficiency descriptions and associated testing procedures to make hundreds of thousands of judgments about examinee language performance. Furthermore, these judgments have generally been considered—with a fairly substantial amount of empirical support—to

provide sufficiently reliable and valid indications of language performance to properly accomplish the practical assessment purpos- for which they were designed and used within the government community. In this regard, development and use of the ILR scale and testing procedure must in all candor be viewed as the most significant and most highly consequential measurement initiative in the proficiency testing field to have occurred over the last three decades.

Nevertheless, despite the high degree of practical utility and the high level of "user satisfaction" that are quite properly attributed to the ILR testing approach, close examination of the ILR level descriptions with regard to the specific psychometric properties that a true measurement scale is expected to demonstrate presents some technical, and ultimately practical, concerns. More explicitly stated, although the fact that individuals can be trained to make reliable judgments that are also viewed as valid by their peers may provide evidence for attibuting scalar properties to the judging system being used, such evidence is not in and of itself sufficient to fully establish the scalar nature of the system: Unidimensionality and other definitional characteristics of a true scale must also be examined and verified.

The unidimensionality question is especially pertinent in the area of human performance. In practice, many scales violate the unidimensionality principle, and with relative impunity. In fact, to the extent that the various component dimensions of a particular construct (such as "language proficiency") are conceptually relatable and are all equally scalable, it is often meaningful and psychometrically appropriate to combine them into a single "scale" and to make judgments based on the values obtained on that single index. This is indeed the approach adopted in defining and using many social-psychological scales, and it is also the implicit, if not always explicit, assumption underlying ILR scale-based proficiency assessment. Note, for example, the original FSI "factor" subscales of accent,

lexicon, structure, fluency, and listening comprehension, each of which was considered to contribute, in a generally additive manner, to the total proficiency rating; as well as the "functional trisection" of content, function, and accuracy, as initially propounded by Higgs and Clifford (1982).

Closer examination of the linguistic aspects underlying the factor subscales and the functional trisection, as well as the observed behavior of both types of subscales as they are used in practice, suggests that both are somewhat problematic in terms of scalability. The factor scores are manifestly not linearly related throughout the ILR scale range in that their proportional contribution to the overall proficiency level rating varies as a function of the level itself, as manifested in the well-known "butterfly" graph (see Fig. 2.1, next page) presented and discussed by Higgs and Clifford (1982). By the same token, the trisection components do not appear to possess equal scalar characteristics. The linguistic *accuracy* of an examinee's performance may reasonably be viewed as lying somewhere along a continuum ranging from none at all to that associated with native-speaker competence. Also, in terms of sociolinguistic *functions,* it could be posited that at least some types of functions can be ordered along a continuum. For example, simple description may be viewed as less demanding than supporting one's opinions or convincing others to change theirs. However, the assumption that the content of a given performance can be similarly ordered is considerably more difficult to support, particularly in view of the fact that in testing practice, the ability to function in specified content domains has been assumed to be both a requirement for a given skill-level rating and evidence that an examinee may deserve it. Recognition of the so-called "hothouse special" phenomenon (i.e., an individual inordinately well versed in a particular lexical domain or specific limited area of discourse) is an acknowledgment of this problem, but only from one point of view—that of requiring the individual so labeled

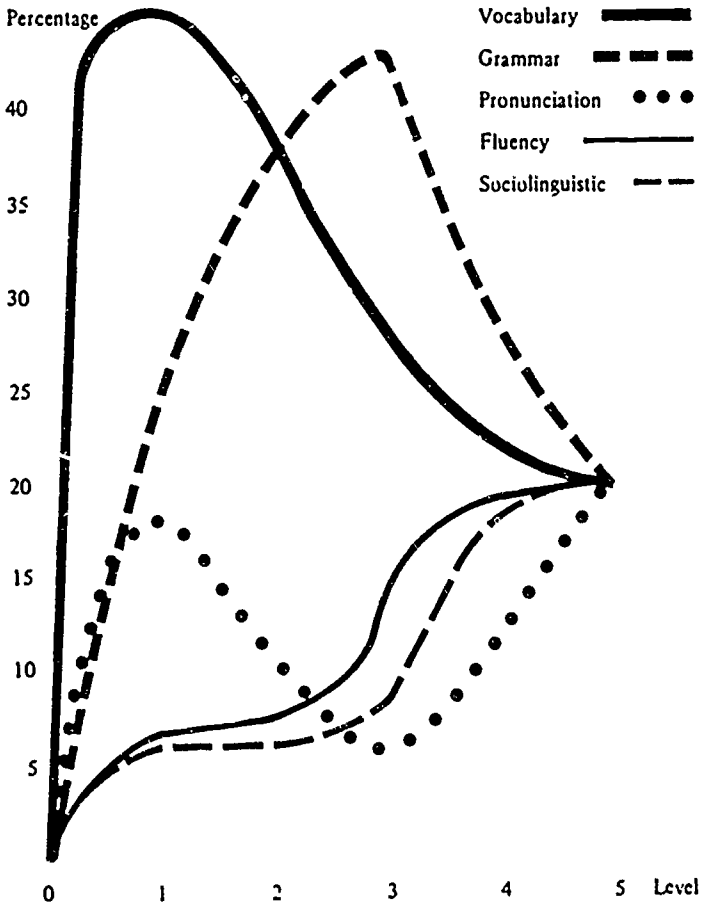


Figure 2.1. Hypothesized Relative Contribution Model

Note. From "The Push Toward Communication" (p. 69) by T.V. Higgs & R.T. Clifford in T.V. Higgs (Ed.), *Curriculum, Competence, and the Foreign Language Teacher.* 1982, Lincolnwood, IL: National Textbook Co. Copyright 1982 by National Textbook Co. Reprinted by permission.

---

to demonstrate control over the other content areas specified in the given level description in order to receive that rating. Whether certain content areas are

intrinsically more difficult to handle than others—a necessary condition for placing them along a difficulty continuum—has yet to be established.

A second area in need of detailed research is that of *across-language* variation in the scaling of particular elements within the accuracy, function, and content domains. For example, with regard to accuracy, the ability to properly convey time/tense is probably a much easier accomplishment (and hence, presumably, lower on the scale) in Chinese than in German. Similarly, expressing apologies would presumably fall at an appreciably different point on the sociolinguistic function continuum for Japanese than for Spanish. Although it would certainly be possible to specify different configurations of the function, content, and accuracy scales for each different language or language group, such an approach would raise serious questions as to the underlying measurement theory or concepts involved, since it would represent an *ad hoc* process of proficiency scale definition rather than an *a priori,* principled, and, most important from the validity standpoint, generalizable approach.

One way to resolve the content dilemma might be to examine it in terms of two separate dimensions of lexicon: *range* of vocabulary and *precision* of vocabulary. Viewing examinee control of "content" in this manner might help reduce the size and significance of the test-versus-real-world sampling problem posed earlier. With regard first to range of vocabulary, a given level of proficiency could be defined as requiring (among other things) the ability to make use of the lexicon associated with a number of areas of discourse. Each higher proficiency level would require control of lexicon in an increasing number of areas. Precision of vocabulary would be scaled on a continuum ranging from virtually no ability to retrieve and use basic lexicon of the content areas in question to a high level of accuracy and sophistication in lexical choice, including the ability to readily express fine nuances of meaning within each content area. However, this approach

would still leave unaddressed the variation in content attributable more to functions than to lexicon.

Another alternative to developing a uniform, hierarchical scale of "content" would be to make no scalar claims whatsoever for the content dimension, but simply to describe the examinee's ability to function within each of several explicitly stated content areas. This approach could give rise to an assessment procedure in which, for example, an examinee would be evaluated as a Level 3 speaker in the areas of politics, fine arts, and economics (i.e., would have satisfied Level 3 function and accuracy requirements within these lexical domains), with no specific performance claims being made or implied for discourse areas outside these particular domains.

One experimental approach to resolving, or at least considerably clarifying, the content issue would be to carry out a concurrent validation study in which given examinees would be administered both a traditional interview and several much more highly specialized "lexical domain" tests in such diverse areas as food, lodging, transportation, autobiographical information, current events, sports, economics, politics, science and technology, and other appropriate specialized topics. At issue would be both the uniformity of a given examinee's performance across the specific lexical domains and the extent to which the traditional interview would be able to predict performance for some or all of the lexical domains. Results of such a study would help to determine (a) whether the traditional interview is doing a reasonable job of standing in for more detailed assessments of examinee control across a variety of content domains; (b) whether and in what respects caution should be exercised in interpreting interview results as reflective of broad "content control"; (c) whether relatively modest changes in the interviewing process with respect to the types of content areas broached might provide appreciable improvements in predictive validity; or (d) whether the correspondence between interview performance and lexical/content

control across a variety of areas is so tenuous as to indicate a need for complete rethinking of the assessment procedure necessary to this task. Experience with the AEI procedure to date suggests that the actual outcomes of the proposed study would involve some combination of the first three possible conclusions; in any event, a considerable amount of properly obtained and carefully analyzed empirical data would be needed to adequately address these issues.

Levels of Measurement

In addition to the question of dimensionality raised in the preceding section, it is necessary to consider the statistical level of measurement (categorical, ordinal, interval, or ratio) represented by the AEI scale. The level-of-measurement question is important because it relates to the types of uses to which the measurement procedure can legitimately be put, as well as to the kinds of statements or inferences that can properly be drawn on the basis of the testing results. A categorical scale would be quite satisfactory for separating individuals into groups, such as those who can do a given job and those who cannot; or those who are better suited for job A than for job B. In both of these cases, a categorical scale is adequate only when differences in *types* of linguistic performance are being evaluated, not greater or lesser *degrees* of performance.

For most uses to which the AEI interview is typically put, an ordinal or, in some instances, interval scale would be required. For example, the task of placing incoming students into appropriate language courses, or that of selecting the best-qualified individuals for important positions requiring language skills, would require at least an ordinal scale, since these applications all involve ranking or ordering the individuals in question

in the language. An even higher level of measurement —the interval scale—is needed when the measurement intent is, for example, to determine how nearly the student has approached the next higher proficiency level since the last assessment was made. Experimental studies involving the use of parametric statistics (e.g., correlational analyses using language proficiency scores as criterion variables in language-learning research) also require interval-level data. The use of noninterval data in such analyses places the results in jeopardy to the extent that the statistical procedure used is affected by violation of the assumption that interval data are in fact being used.

Although considerable evidence exists, both statistical and anecdotal, that the AEI and similar proficiency scales are at least ordinal in nature, the question of whether they are interval as well is more complex because of the empirical difficulties involved in investigating certain fundamental assumptions associated with language proficiency theory as embodied in the AEI guidelines and testing approach. For example, it is generally presumed, and indeed asserted, by those involved in AEI-type testing that the "distance" between any two levels increases as one progresses up the proficiency scale, and that the "plus" point within any given level is closer to the value of the next higher level, rather than midway between the two levels. Although certified testers are taught to make judgments based on these assumptions, to determine whether the scale properties implicit in these assumptions have any basis in external "fact" would require that the proficiency level ratings themselves be compared against some other scale of known interval properties and of such a nature as to have a high degree of face validity within the proficiency testing movement. Various investigations that might be carried out within the overall research framework described in the following section could help to resolve this question, as well as others posed earlier in this chapter.

A Proposed Framework for Language Proficiency Testing Research

Bachman and Clark (1987) proposed the detailed and systematic investigation of a wide variety of proficiency-based testing issues within the framework of a single large-scale research program, under whose auspices a number of individuals and institutions would be asked to combine their efforts in an attempt to provide reliable and detailed information on each of these issues. In outline form, the major steps in the proposed program are to:

1. specify a prototypical model of communicative language ability, including but not limited to the performance elements currently at issue in the AEI context;
2. develop highly detailed and comprehensive performance tests covering each of the component aspects of the communicative ability model;
3. develop shorter, necessarily less comprehensive, practically oriented tests that could be used in real-life applications as effective surrogates for the criterion measures at issue in (2);
4. establish the types of construct validation data to be gathered to support or contradict the validity of the communicative ability model, the criterion measures, and the practically oriented tests;
5. carry out a series of research studies in which the criterion performance test, practically oriented tests, nd construct validation data are intercorrelated and ompared on a two- and three-way basis;
6. draw necessary conclusions based on these studies and correspondingly revise the proficiency model, criterion tests, and/or practically oriented tests;
7. repeat steps (5) and (6) until the researchers involved express high confidence in the theoretical and psychometric quality of the data, and operational test users find high practical utility in the results.

A comprehensive study conducted along the general lines outlined here would fully accommodate the investigation of the various issues raised in this chapter concerning the AEI scale and testing procedure. In addition, it would do so within the context of an expanded communicative ability model that might offer some useful new perspectives for the further elaboration of the AEI approach and/or for the development of a variety of other types of instruments and procedures in the service of developing increasingly valid, reliable, and practical language proficiency testing.

References

Adams, M.L. (1978). Measuring foreign language speaking proficiency: A study of agreement among raters. In J.L.D. Clark (Ed.), *Direct testing of speaking proficiency: Theory and application.* Princeton, NJ: Educational Testing Service.

Bachman, L.F., & Clark, J.L.D. (1987). The measurement of foreign/second language proficiency. *Annals of the American Academy of Political and Social Science, 490* (March), 20-33.

Burt, M.K. (1975). Error analysis in the adult EFL classroom. *TESOL Quarterly, 9,* 53-63.

Carroll, J.B., Clark, J.L.D., Goddu, R.J.B., Edwards, T.M., & Handrick, F.A. (1966). *A parametric study of language training in the Peace Corps.* Cambridge, MA: Laboratory for Research in Instruction, Harvard Graduate School of Education.

Chastain, K. (1980). Native speaker reaction to instructor-identified student second-language errors. *Modern Language Journal, 64,* 210-15.

Clark, J.L.D. (1978). Interview testing research at Educational Testing Service. In J.L.D. Clark (Ed.), *Direct testing of speaking proficiency: Theory and application.* Princeton, NJ: Educational Testing Service.

Clark, J.L.D. (1979). Direct vs. semi-direct tests of speaking ability. In E.J. Briere & F.B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies.* Washington, DC: Teachers of English to Speakers of Other Languages.

Clark, J.L.D. (1981). Survey measures: Language; Survey measures: Results. In T.S. Barrows (Ed.), *College students' knowledge and beliefs: A survey of global understanding.* New Rochelle, NY: Change Magazine Press.

Clark, J.L.D. (1986). Development of a tape-mediated, ACTFL/ILR scale-based test of Chinese speaking proficiency. In C.W. Stansfield (Ed.), *Technology and language testing.* Washington, DC: Teachers of English to Speakers of Other Languages.

Clark, J.L.D., & Li, Y.C. (1986). *Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages* (Final project report for Grant No. G008402258, U.S. Department of Education). Washington, DC: Center for Applied Linguistics.

Experiment in International Living. (1976). *Your objectives, guidelines, and assessment: An evaluation form of communicative competence* (rev. ed.). Brattleboro, VT: Author.

Galloway, V.B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64,* 428-433.

Higgs, T.V., & Clifford, R.T. (1982). The push toward communication. In T.V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher* (ACTFL Foreign Language Education Series, Vol. 13). Skokie, IL: National Textbook Co.

Ludwig, J. (1982). Native-speaker judgments of second language learners' efforts at communication: A review. *Modern Language Journal, 66*, 274-283.

Mattran, K.J. (1977). Native speaker reactions to speakers of ESL. *TESOL Quarterly, 11,* 407-414.

Munby, J. (1978). *Communicative syllabus design: A sociolinguistic model for defining the content of purpose-specific language programmes.* Cambridge, England: Cambridge University Press.

Raisler, I. (1976). Differential response to the same message delivered by native and foreign speakers. *Foreign Language Annals, 9,* 256-259.

Shohamy, E. (1983). Rater reliability of the oral interview speaking test. *Foreign Language Annals, 16,* 219-222.

III

Issues Concerning the Less Commonly Taught Languages

by Irene Thompson, George Washington University;
Richard T. Thompson, ACTFL/
Center for Applied Linguistics;
and David Hiple, ACTFL

This chapter addresses the application of proficiency guidelines to the less commonly taught languages. Questions of relevance and appropriateness in theory and practice will be addressed. A distinction is drawn among commonly, less commonly, and much less commonly taught languages. This distinction is more practical than theoretical and relates to questions of supply and demand, the need for priority-setting, the availability of trained specialists in specific languages as well as the likelihood of developing such specialists in many of these languages.

Questions of Eurocentric bias and the impact of the application of the provisional generic guidelines to languages with different typologies, such as Chinese, Japanese, and Arabic, and their role in a subsequent redefinition of the generic guidelines themselves are traced for both speaking and reading.

Theoretical and practical problems in adapting the guidelines to specific less commonly taught languages are discussed, including the presence of Hindi-English code switching at high levels of proficiency in educated native speakers, special problems of diglossia in Arabic, complex inflectional morphologies in languages such

as Russian, the early appearance of significant problems in register in Indonesian and Japanese, and the complex and predominantly nonphonologically based writing systems of languages such as Chinese and Japanese with repercussions for the generic reading and writing guidelines.

Finally, policy issues affecting the various constituencies are addressed, including the role of the federal government, the language and area studies centers most directly affected by recent federal legislation, and pending regulations relating to proficiency testing and competency-based language programs.

Foreign Language Enrollments

A survey by Brod and Devens (1985) indicates that in 1983, a total of 784,515 college students were enrolled in Spanish, French, and German language courses as follows: 386,238 in Spanish, 270,123 in French, and 128,154 in German. These three languages have come to be known as the commonly taught languages.

Enrollment figures in 1983 reveal a second cluster of foreign languages taught at the college level (following Spanish, French, and German). These were Italian, 38,672; Russian, 30,386; Hebrew, 18,199; Japanese, 16,127; Chinese, 13,178; Portuguese, 4,447; and Arabic, 3,436. To illustrate the comparative significance of these figures, the number of students enrolled in Japanese courses in 1983 represented about 4 percent of the number of students enrolled in Spanish courses, and the number students enrolled in Arabic courses represented about 1 percent of those enrolled in Spanish. This second cluster of languages is often referred to as the less commonly taught languages.

After this cluster of less commonly taught languages, enrollments reveal that most other foreign languages are much less commonly taught. In 1983, for

example, 507 college students enrolled in Swahili courses, 219 college students studied Hindi, and 85 students enrolled in Indonesian courses, the latter figure representing approximately 2 percent of enrollments in Arabic. Yet, even Indonesian scholars, with their 85 students, could take comfort in the fact that only 14 students enrolled in Uzbek, while only 4 students enrolled in Ibo. These languages may thus be referred to as the much less commonly taught languages.

At the secondary level, differences between enrollments in *commonly taught* and *less commonly taught* languages are even more dramatic. An American Council on the Teaching of Foreign Languages (ACTFL) survey (1984) indicated that Spanish, French, German, Latin, and Italian, in that order, accounted for approximately 99 percent of the 2,740,198 foreign language enrollments. Of the remaining 1 percent, 5,497 students were enrolled in Russian, 1,980 in Chinese, and 51 in Arabic.

Thus, within what has been referred to as the *less commonly taught languages,* there is a wide range in student enrollments, reflecting the fact that some languages are, indeed, *much less commonly taught.* This distribution provides a basis for priority-setting in the face of limited training resources, both human and financial.

ACTFL Proficiency Initiatives Beyond the Commonly Taught Languages

ACTFL's activities in extending the language proficiency assessment movement beyond government and into academia have progressed from projects in *commonly taught languages,* initiated in 1981, to projects in *less commonly taught languages,* begun in 1983, to projects in *much less commonly taught languages,* started in 1985. The initial projects involved the

development of proficiency guidelines for French, German, and Spanish, as well as the training of individuals to administer and evaluate oral proficiency tests in Spanish, French, German, and Italian. The second stage of activities involved developing proficiency guidelines for Chinese, Japanese, and Russian, and oral proficiency tester training in Arabic, Chinese, ESL/EFL, Japanese, Portuguese, and Russian. The third stage of activities involved a dissemination project, undertaken by ACTFL jointly with the Center for Applied Linguistics, to extend proficiency concepts to Hindi, Indonesian, and Swahili, as well as preliminary oral proficiency tester-training activities in Hindi, Indonesian, Swahili, and a small sample of other African languages such as Hausa and Lingala.

Development of Proficiency Guidelines for the Less Commonly Taught Languages

In 1983, ACTFL received support from the U.S. Department of Education to initiate the second stage of the guidelines project to create language-specific proficiency statements for Chinese, Japanese, and Russian. As the working committees began their task, a West European bias of the existing generic guidelines became most evident in statements concerning accuracy in speaking and in statements dealing with the writing system (Hiple, 1987).

With respect to accuracy in speaking, this bias was seen in references to grammatical constructions common to most West European languages, such as inflections, subject-verb and adjective-noun agreement, articles, prepositions, tense usage, passive constructions, question forms, and relative clauses. It was felt that learners of languages such as Japanese, Chinese, and Arabic had to deal with a very different set of grammatical constructions, and thus that the accuracy statements in the generic guidelines were not relevant to

these languages. To some extent, this was true even of Russian, which, for instance, lacks articles.

After completing the initial draft of the Chinese, Japanese, and Russian guidelines, it became evident that creating meaningful guidelines for those languages would not be possible without a revision of the generic guidelines. As a result, ACTFL petitioned the Department of Education for and was granted an amendment to the project to revise the generic guidelines in order to broaden them enough to accommodate language-specific statements for Chinese, Japanese, and Russian.

With respect to writing, the West European bias of the generic guidelines was even more obvious. The guidelines assumed that learners would not require much time to master the principles and mechanics of the writing system. Therefore, Novice writers were assumed to be capable of functions such as filling out simple forms and making simple lists. This was felt to be unrepresentative of languages such as Japanese, Chinese, and Arabic, in which the complexities of the writing system dictated a more graduated set of statements regarding the development of real-life writing functions.

The evolution of the proficiency guidelines as related to less commonly taught languages can be traced from the Interagency Language Roundtable (ILR) definitions through the Provisional Proficiency Guidelines to the revised Proficiency Guidelines and the respective language-specific descriptions in several uncommonly taught languages. Space does not allow inclusion of all the changes in the definitions for all levels in all four skills, and only selected levels in speaking and reading are discussed here.

Evolution of the Speaking Guidelines. The ILR definition for Level 1 speaking proficiency (S-1) and the corresponding ACTFL provisional definitions for the Intermediate-Low and Intermediate-Mid levels serve to illustrate the evolution of the speaking guidelines. It

will be remembered that S-1 (Elementary Proficiency) is defined as being able to "satisfy minimum courtesy requirements and maintain very simple face-to-face conversations on familiar topics."

An obvious difference between the ILR S-1 definition and the ACTFL Intermediate descriptions is that the equivalent to the ILR S-1 is represented by two subranges—Intermediate-Low and Intermediate-Mid—on the ACTFL scale. Liskin-Gasparro (1984) describes the Common Yardstick Project of the Educational Testing Service (ETS) and the need for a scale that discriminates more finely at the lower end, where most of the foreign language students in schools and colleges tend to cluster. This need is particularly real in less commonly taught languages, whose students can expect to invest more time in order to arrive at the Intermediate level than students of commonly taught languages. For example, the School of Language Studies of the Foreign Service Institute estimates that students may require twice as much time to attain S-1 proficiency in Arabic, Chinese, Japanese, and Korean as to attain the same level of proficiency in Spanish or French. Thus, the need to distinguish among subranges of the Intermediate level of proficiency seems particularly compelling for a number of less commonly taught languages.

Content/context. The most noticeable aspect of the ILR S-1 definition is its orientation toward satisfying minimum courtesy requirements, survival needs such as getting food and lodging, and work demands such as giving information about business hours and explaining routine procedures. ACTFL's Provisional Guidelines retained the courtesy and survival requirements but left out reference to specific work demands. Instead, the requirement of satisfying limited social demands was added at the Intermediate-Mid level. The language-specific Provisional Guidelines began the process of adaptation of content to the academic environment by including contexts appropriate for academic learners, such as references to school (French and Spanish

Intermediate-Low), learning the target language and other academic studies (German Intermediate-Low), autobiographical information, leisure time activities, daily schedule, future plans (French and Spanish Intermediate-Mid), and academic subjects (German Intermediate-Mid).

This process of content/context adaptation continued in the 1986 version of the ACTFL guidelines with the introduction at the Intermediate-Low level of the more general statement "Able to handle successfully a limited number of interactive, task-oriented and social situations." As a result, language-specific statements in the revised language-specific guidelines of 1986 include references to greetings, introductions, simple biographical information, social amenities, making, accepting, or declining invitations, handling routine exchanges with authorities, and making social arrangements.

Accuracy. When it came to accuracy, the problem of adaptation was more serious, because many of the accuracy requirements in the ILR descriptions and the ACTFL Provisional Guidelines were typically reflective of Indo-European languages and irrelevant for Arabic, Chinese, Japanese, and Russian.

An attempt was made, therefore, to remove many of the quality statements and reserve them for their proper place in the language-specific guidelines. For instance, the Intermediate-Mid description in the Provisional Guidelines contained references to subject-verb agreement, adjective-noun agreement, and inflections. In the revised guidelines of 1986, all references to these structures were removed. As a result of this revision, the language-specific guidelines were free to include accuracy statements that were more representative of their languages. Thus, the Arabic guidelines specify verb-object phrases, common adverbials, word order, and negation (Allen, 1984); the Chinese guidelines refer to word order, auxiliaries and time markers (ACTFL, 1986a); the Japanese guidelines single out formal nonpast/past, affirmative/negative forms,

demonstratives, classifiers, and particles (ACTFL, 1987); and the Russian guidelines include references to adjective-noun and subject-predicate agreement, and a developmental hierarchy of cases (ACTFL, 1986c).

A related refinement was the reformatting of the guidelines to present the generic and the language-specific statements together so that the two could be viewed simultaneously. This change in format was particularly useful in light of the attempt to maintain the neutrality of the generic descriptions and to focus on the accuracy statements in the language-specific descriptions. This proved helpful in oral proficiency tester training workshops because it was no longer necessary to deal with two separate documents. Also, language-specific guidelines no longer needed to repeat the same generic statements and could concentrate instead on providing appropriate language-specific examples.

Evolution of the Reading Guidelines. The most noticeable aspect of the reading guidelines for the ACTFL Novice-Low and Novice-Mid (ILR R-0) levels is complete negativity—"Consistently misunderstands or cannot comprehend at all." In the Provisional Guidelines, only the Novice-Low level was characterized negatively as "No functional ability in reading the foreign language." However, this negative wording was felt to be unhelpful, because it focused excessively on what the candidates could not do, and not on what they could do in the target language. Therefore, the revised generic statements for reading substituted a positive statement that recognizes the beginning of reading development: "Able occasionally to identify isolated words and/or major phrases when strongly supported by context." This allowed the Chinese Novice-Low description to include a reference to "some Romanization symbols and a few simple characters." At the same time, the Russian examinee at the Novice-Low level recognizes some letters of the Cyrillic alphabet in printed form.

When the ILR R-0 level was divided into two subranges for academic use, the Novice-Mid description

allowed for some development of reading ability by stating: "Sufficient understanding of the written language to interpret highly contextualized words or cognates within predictable areas. Vocabulary for comprehension limited to simple elementary needs, such as names, addresses, dates, street signs, building names, short informative signs (e.g., no smoking, entrance/exit) and formulaic vocabulary requesting same." Although this positive wording was a step in the right direction, reference to cognates and the specificity of the examples posed a number of problems for noncognate languages with nonalphabetic writing systems. Chinese, for example, shares no cognates with English; students of Chinese must learn both characters and Romanization system(s). Thus the reading of names, for instance, becomes a rather advanced skill.

The revised generic guidelines of 1986 distinguish among alphabetic, syllabic, and character-based writing systems, thus allowing greater latitude for languages such as Chinese and Japanese. The reference to cognates was modified as follows: "The reader can identify an increasing number of highly contextualized words and/or phrases including cognates and borrowed words, where appropriate." All references to specific materials representative of this level were left out.

As a result of these changes in the generic guidelines, the Chinese description of the Novice-Mid reader includes the ability to identify/recognize a small set of typeset or carefully hand-printed radicals and characters in traditional full form or in simplified form, and full control over at least one Romanization system. The reading context includes public writing in high-context situations, such as characters for "male" and "female" on restroom doors. In contrast, the Novice-Mid reader in Russian can identify all letters of the Cyrillic alphabet in printed form and can read personal names, street signs, public signs, and some names on maps. The Arabic Novice-Mid reader can identify the letters but has difficulty recognizing all four forms of each letter as well as the way in which these letters are

joined to each other in forming words. He or she can recognize individual Arabic words from memorized lists as well as highly contextualized words and cognates such as public signs.

The question arises whether such changes in the language-specific statements have led to a lack of comparability across sets of guidelines, weakening the unity of the system. This problem resulted from ACTFL's decision to subdivide the lower-level descriptions into three subranges, Novice-Low, Novice-Mid, and Novice-High, so that learner achievement at the very beginning stages of language learning could be recognized. The entire Novice level is prefunctional, and, as such, represents a developmental sequence leading to the first functional level (i.e., the Intermediate level), rather than an index of real-life functional ability. If the description of this developmental sequence is to have any validity, the statements at the Novice level will have to take on a somewhat divergent character. This divergence should not represent serious difficulties for testing, because any test at the Novice level will turn out to be a test of achievement, not of proficiency.

Theoretical and Practical Problems in Adapting the Proficiency Guidelines to Specific Languages

On the practical side, there is little doubt that the proficiency guidelines have succeeded in injecting some vitality into the language teaching field by offering both a framework for program planning and an instrument for assessing student progress.

On the theoretical side, the development of language proficiency guidelines is an ambitious attempt to define and quantify foreign language proficiency at various points in its development and to describe the essential features of each stage in a few well-chosen sentences accompanied by a few carefully selected examples. For such an attempt to be successful, it must be a dynamic

process that involves constant refinement of current understanding of language competency and continual validation of language content and of real-world language-use situations typical of language performance at different levels of proficiency (for a detailed discussion of validation of second-language testing, see Clark & Lett, this volume). Because this formidable task can never be complete, the guidelines will always reflect a stage in the developing understanding of the dynamics of interlanguage.

For the moment, the guidelines have raised many questions that cannot be answered due to lack of empirical research. This lack may be partially due to the relatively recent introduction of the guidelines into the academic setting. The following partial research agenda applies to all languages:

1. validation of the claims in the guidelines regarding the developmental hierarchies of different aspects of linguistic performance including pragmatic and discourse strategies in different languages;
2. determination of differences in specific aspects of linguistic performance across major scale boundaries;
3. examination of specific linguistic features that distinguish planned from unplanned discourse; and
4. examination of the validity of the developmental sequence outlined in the receptive skills guidelines.

In addition to these problems, which affect all languages for which guidelines have been developed or are being developed, the development of guidelines for languages with different typologies has brought forth a host of problems that hitherto had not been dealt with. These problems are examined next in the context of African languages, Arabic, Chinese, Hindi, Indonesian, Japanese, and Russian.

The Case of Russian. The availability of government testers to train the initial contingent of academic testers in Russian made it possible for a group of trained

individuals to begin work on the Russian guidelines in 1984.[1] According to Thompson (1987), unlike the other less commonly taught languages, Russian, an Indo-European language, faced no special challenges in developing language-specific guidelines from a generic starter kit. This adaptation process could be best characterized by a conflict between the desire to make the level descriptions come to life through a variety of examples and the desire to preserve the global character of these descriptions.

The process of adaptation was also not without some uneasiness caused by a conflict between the desire to make the Russian guidelines conform to those in French, German, and Spanish, and the need to include in them references to features that are unique to Russian. These features were related to content/context, accuracy, and the lack of provision in the generic guidelines at the Novice level for the learning of another alphabetical system.

With respect to content/context, the committee members felt that they had to correct the West-European and adult/professional bias in favor of contexts in which U.S. students in the Soviet Union would most likely find themselves, and to provide examples of topics that Americans would most likely discuss with Russians, particularly at the Advanced and Superior levels. It was felt that many of the survival situations mentioned in the French, Spanish, and German guidelines could not be applied to Russian because they would either simply not occur in the Soviet Union or they would be structured differently. This is true even at the Novice level, where some familiar categories in West European languages (e.g., telling time, days of the week, months) require Intermediate-level skills in Russian.

Somewhat more serious problems presented themselves in the area of accuracy. For example, the provisional guidelines in the Advanced-Plus description referred to lack of accuracy as follows: "Areas of weakness range from simple constructions such as plurals, articles, prepositions, and negatives to more complex

structures such as tense usage, passive constructions, word order, and relative clauses." Such specificity caused problems for Russian, in which additional grammatical categories such as pronominal, adjectival and nominal declensions, verbal aspect, modality, verbs of motion, and prefixation, among others, present significant difficulties for the learners. The case of Russian indicated that considerable variation in both inventories of structural features and their hierarchical development had to be accommodated before the Superior level.

The removal of references to specific structures in the revised guidelines of 1986 facilitated the subsequent revision of the Russian-specific guidelines, for the revision committee members[2] no longer felt constrained by the imposition of developmental hierarchies for grammar more characteristic of less inflected languages. As a result, the revised Russian guidelines in the description of the Advanced-Plus speaker refer to cases, aspect, mood, word order, and the use of particles.

Although all members of the Russian guidelines committee had been trained in the administration of the oral proficiency interview and all were experienced teachers of Russian, they were somewhat uneasy about positing a developmental hierarchy of acquisition of grammatical, discourse, sociolinguistic, and pragmatic features on the basis of observation and experience rather than research evidence. It was felt then, and is felt still, that the availability of large amounts of data from taped oral interviews in Russian should provide the impetus for psycholinguistic research into characteristics of learner speech at different levels of proficiency, such as suggested by Byrnes (1987). The results of this research may guide efforts to reexamine and reevaluate some of the statements in the current version of the Russian proficiency guidelines with regard to various aspects of learner performance at different levels of proficiency, keeping in mind, of course, the danger of a cyclical effect in using interview data to validate oral proficiency interview traits.

Finally, the lack of accommodation in the provisional reading/writing guidelines for learning to recognize/produce Cyrillic letters caused some concern. While some transfer can be made from West European languages when it comes to recognizing/producing Cyrillic letters, most of them represent a different pattern of letter-sound correspondence or have different shapes altogether. In addition, the printed and longhand versions of the letters look quite different. Thus, learners must go through a training period before they are able to recognize/produce Cyrillic script. The provisional guidelines, however, describe the Novice-Low reader as having no functional reading ability, while the Novice-Mid reader is described as already able to read highly contextualized words or cognates within predictable areas such as names, addresses, dates, street signs, building names, short informative signs, and so on. It was felt that there was a discontinuity between these two subranges that did not reflect the early stages of learning to read in Russian.

The problem was solved by having the Novice-Low reader in Russian recognize some letters of the Cyrillic alphabet in printed form and a few international words and names. By contrast, the Novice-Mid reader could identify all letters of the Cyrillic alphabet in printed form and some contextualized words such as names, public signs, and so on. Finally, the Novice-High reader could identify various typefaces in printed form or in longhand as well as highly contextualized words, phrases, and sentences on maps and buildings and in schedules, documents, newspapers, and simple personal notes. In this manner, the Novice level was designed to represent the gradual beginning steps in learning to read Russian.

The recommendations of the National Committee on Russian Language Study (AAASS, 1983), which called for the development of a common metric and its use to set standards for Russian language study, helped pave the way for the introduction of the proficiency guidelines and the oral proficiency interview into the Russian

field. The response has been generally quite positive, and during the past few years there has been a good deal of activity in the field involving the guidelines and the oral interview test. The following deserve mention: (a) there is now a contingent of more than a dozen certified oral proficiency testers and several academic tester-trainers in Russian, thereby ending dependence on the U.S. government for training; (b) curriculum workshops geared to teaching for proficiency are now being offered to secondary and postsecondary Russian language teachers throughout the academic year and especially during summers at various locations nationwide; (c) videos of oral interview tests at all levels were developed at Middlebury College under a grant from the Social Science Research Council for use in tester training; (d) the Educational Testing Service has developed Advanced Russian Listening and Reading test based on ACTFL Listening and Reading Guidelines that reports raw scores and/or proficiency ratings from Intermediate-High (ILR 1+) to Superior (ILR 3) or higher; (e) major Russian overseas programs such as the Council on International Educational Exchange, the American Council of Teachers of Russian, and the Middlebury College Russian Program at the Pushkin Institute in Moscow use the ACTFL oral proficiency interview and the ETS Advanced Listening/Reading Test for pre- and post-program evaluation of participants, and data are being collected to update Carroll's (1967) study with respect to Russian; and (f) some institutions have introduced graduation requirements for undergraduate and graduate majors in Russian in terms of proficiency levels in various skills. Other institutions are using the oral proficiency interview to screen prospective teaching assistants. Many institutions are reevaluating their language courses by setting objectives in terms of proficiency levels in various skill combinations.

The Case of Hindi. Hindi presents another set of problems hitherto not encountered in the development

of guidelines for other languages. Specifically, accommodations must be made for Hindi-English code switching. Code switching is generally an indication of a relatively low level of proficiency in a second language, but in the case of Hindi, appropriate Hindi-English code switching is representative of educated native Hindi speakers.

Secondly, in terms of extending the oral proficiency interview to additional less commonly taught languages, a major problem presented itself when no government tester was available to train academic testers in a particular language. ACTFL first addressed this problem in the case of Hindi. The solution, a time-consuming one, was to train testers in a language other than the target language, and then to help the most interested ones transfer the concepts, procedures, and rating criteria to the target language.

Several problems had to be resolved before guidelines could be created for Hindi (Gambhir, 1987). According to Gambhir, it was desirable to study the concept of an educated native speaker in the multilingual speech community of India, where English is used by the educated elite in most formal and professional domains, and where Hindi is primarily relegated to the more restricted domain of informal socialization and areas of higher education dealing with language, literature, and culture. Because of the widespread use of English, which is the co-official language of India along with Hindi, in government, education, science, technology, and commerce, most educated native speakers of Hindi lack opportunities to develop higher levels of proficiency normally associated with professional, educational, and formal domains of language use. If the Hindi guidelines are to reflect the actual use of Hindi by educated native speakers, these limitations must be taken into account.

According to Gambhir (1987), an additional problem is the presence of two styles in the speech of educated native speakers of Hindi. The spoken style, which contains many borrowings from English, Persian, and

Arabic, is used in speech and writing for informal purposes, while the written style, which contains many Sanskrit words, is reserved for formal speech and writing. The spoken style is characterized by frequent, rule-governed, Hindi-English code switching when used with Hindi-English bilinguals, which does not occur in interactions between Hindi-English bilinguals and monolingual speakers of Hindi. Since educated native speakers of Hindi use both a mixed and an unmixed code in informal speech and writing, the proficiency guidelines for Hindi must reflect this aspect of sociolinguistic competence. The situation is paradoxical: Hindi-English code switching indicates lesser proficiency in everyday, survival situations but greater proficiency in formal, professional settings.

In addition, the content/context in which Hindi is used needs to be elucidated. For instance, according to Gambhir, the Superior-level functions in Hindi are mostly exercised in the areas of language, culture, and literature; the Advanced-level functions occur mostly in informal social situations; and the Intermediate-level functions occur mostly in rural areas and in contacts with uneducated or less educated native speakers of Hindi who normally have little or no contact with foreigners.

Finally, in addition to making accuracy statements regarding control of various phonological, morphosyntactic, and discourse features of Hindi, statements regarding sociolinguistic competence must take into account the complexity of rules governing style according to the relative status, age, sex, and relationship of the interlocutors as well as the formality/informality of the situation.

To decide which linguistic features should be expected to be fully, partially, or conceptually controlled at which level, Gambhir suggests combining two different approaches—one based on experience as to what to expect in terms of functions, content/context, and accuracy at what level, and the other based on an analysis of a large number of interviews at different

levels. This combined approach requires a tentative formulation of level descriptions through analysis of actual interviews and supplementing missing data with observations based on experience.

The process is already under way. Hindi testers[3] trained in administering the oral proficiency interview in ESL started the transfer by administering the interview in Hindi. They identified the best Hindi interviews, translated the questions asked in those interviews into English, and met with experienced ESL testers to obtain feedback on elicitation techniques and assistance in rating the samples.

In late spring of 1987, an oral proficiency testing workshop was conducted for professors of Hindi and Tamil at the South Asia Department of the University of California at Berkeley, with representatives from Columbia University and the University of Washington also attending. Half the practice interviews were conducted in Hindi, and half were conducted in English. The Hindi testers are working toward tester certification in Hindi, and the Tamil testers are working through English initially, with the long-term goal of becoming testers in Tamil. A second tester-training workshop in Hindi has been scheduled in April 1988 at the University of Illinois at Urbana.

Vijay Gambhir and the University of Pennsylvania have won a Department of Education grant to create proficiency guidelines for Hindi. The development of guidelines and the training of testers will be mutually reinforcing, because experiential knowledge acquired through testing will contribute to the writing of guidelines, and the guidelines will ultimately improve testers' ability to elicit and reliably rate speech samples.

The Case of Indonesian. Wolff (1987) reported that he attended a German tester-training workshop and then conducted about 20 half-hour interviews ranging from Novice to Superior with students of Indonesian at Cornell University. The interviews were presented to a small group of individuals at the 1986 meeting of the

Association for Asian Studies and discussed with oral proficiency interview experts from ACTFL and the Central Intelligence Agency with regard to their content and as to what they showed about the basic characteristics of students at different levels.

As a result of this preliminary work, Wolff does not think that there are any particular features of Indonesian that could not be measured by a common metric, even though Indonesian is significantly different from the more commonly taught West European languages. Despite the fact that the grammar of Indonesian is based on a totally different set of principles than those on which most commonly taught Indo-European languages are based, there is no reason, according to Wolff, why the generic guidelines expressed in terms of functional abilities at different levels would not be applicable to Indonesian as well.

Wolff suggests that the next step in the development of guidelines for Indonesian is the determination of the features of phonology, grammar, and vocabulary that can be associated with each stage of proficiency in Indonesian. In addition, Wolff thinks that a determination should be made as to the candidate's ability to make use of the appropriate style, register, and sociolinguistic rules of Indonesian. Wolff makes the point that these rules are quite rigid, and that Indonesians do not have a great amount of tolerance for deviation from sociolinguistic norms. Even the simplest utterance must adhere to rules for acknowledging the relative social status of the conversational partners through appropriate use of various sociolinguistic rules, such as those for forms of address. As a result, a set of guidelines for testing students' communicative ability in Indonesian will have to include specific statements regarding degree of control of these sociolinguistic rules at different levels of proficiency. Wolff makes the additional point that oral proficiency interviews in Indonesian must include routines that require the use of sociolinguistic rules representative of various settings peculiar to the Indonesian culture.

In terms of actual need, Wolff thinks that the development of proficiency guidelines for Indonesian is a worthwhile endeavor for two reasons. First, Indonesia is the fifth largest country in the world and is one of the few former colonies in which the local language has truly become a national language. Secondly, although the total number of people studying Indonesian is not very large, they represent different levels of proficiency, and it is important to be able to assess their competence.

As Wolff sees it, however, a number of problems arise when it comes to testing a significant number of these individuals. These stem from the fact that Indonesian is largely taught by the linguist-native informant method. Linguists often have too many other professional responsibilities to devote much time to proficiency testing, and native informants are typically temporarily employed and are not professional language teachers. Wolff suggests as one possible solution the development of a semi-direct test of speaking proficiency that would be validated against the ACTFL interview. Experience with a semi-direct test of speaking proficiency for Chinese, reflecting as closely as possible the functions and content of the oral proficiency interview, shows that the former correlates highly with the latter (Clark, 1986). This would not eliminate the need for tester training, however, because trained specialists would still be required to evaluate the taped material. The chief advantage of a semi-direct over a direct test is the elimination of travel time and expense for face-to-face contact with the examinee. This is a major consideration in the case of a less commonly taught language such as Indonesian, given the sparse, widely separated examinees and the difficulty in maintaining the skills of testers who would not be engaged in proficiency testing on a regular basis. The main disadvantage of a semi-direct approach is that oral responses to tape-recorded or printed stimulus material do not allow examinees to demonstrate their ability to use discourse strategies (Clark & Lett, this volume). An alternative solution would be to adopt the government

practice of conducting interviews with two testers—a native speaker trained to elicit a ratable speech sample, and a linguist experienced in both evaluation and elicitation techniques. This solution would require availability of testing teams throughout the country.

In late spring 1987, two additional professors of Indonesian, one each from the University of Wisconsin and Ohio State University, trained as oral proficiency testers in English as a first step toward becoming testers in Indonesian. These testers will seek certification in English and then begin to test experimentally in Indonesian to gather information about the application of the oral interview to that language.

The Case of Arabic. In the 1980s, as Arabists began preparing to introduce proficiency testing to their field and to develop proficiency guidelines in Arabic, they had to (a) define the spectrum of language use that characterizes an "educated native speaker of Arabic"; (b) identify the registers of language that such a person habitually uses and the processes of switching between them; and (c) determine the impact of all these factors on oral proficiency testing and proficiency guidelines. Allen (1984, 1987) explains the problem as follows. The language used for oral communication in a given Arab community is the language that people learn at home, and it can be one of a number of colloquial dialects that vary from country to country and from one community to another. Geographically contiguous colloquial dialects are mutually comprehensible, but geographically separated ones are less so. As a result, in certain situations, a form of the standard written language, referred to as Modern Standard Arabic (MSA), which Arabs learn in school, is used for oral communication. Thus, the colloquial dialect is reserved for day-to-day usage, while MSA is generally restricted to formal situations such as lectures, newscasts, and pan-Arab and international conferences. MSA, being a literary language, is almost exclusively used for all writing purposes, formal and informal. The major exceptions

are some Egyptian drama and any dialectal folk poetry.

This diglossic situation creates a major problem in proficiency testing. According to McCarus (1987), academic programs in the United States generally teach MSA because very few programs can afford to teach one or more dialects as well. The result is a somewhat anomalous situation in which a student might be an Intermediate-High or even an Advanced speaker when it comes to formal or academic discussion involving politics or religion, for example, but only a Novice-High or Intermediate-Low when it comes to dealing with basic survival situations.

Two solutions have been proposed. One is to ignore the dialects and write the guidelines for MSA alone. The other is to choose, in addition to MSA, a major colloquial dialect, such as Egyptian, and write two sets of guidelines, one for MSA and one for the dialect.

Allen and his associates at the University of Pennsylvania have received a Department of Education grant to prepare a set of provisional proficiency guidelines for Arabic. After considering the advantages and disadvantages of the two solutions, the committee decided to write one set of guidelines. As a result of their work, a preliminary set of Arabic Guidelines was published in *Al-ᶜArabiyya* (Allen, 1985). This set has been renamed the Provisional Guidelines for Modern Standard Arabic (as opposed to Arabic alone).

The next step, according to Allen (1987), is to set up a measure of appropriateness of language usage and rate the candidate's ability to use colloquial Arabic or MSA according to sociolinguistic rules adhered to by native speakers of Arabic. McCarus (personal communication) also supports the idea of a single set of guidelines for Arabic. He points out that if the proficiency levels for each of the four language skills are defined, it is up to the individual to do the best he or she can, using colloquial Arabic or MSA, where appropriate. The examiner should represent the Arab region for which the person is being tested (e.g., an Egyptian tester if the examinee is going to Egypt). In other cases, for a general rating in

Arabic, the tester and the examinee should control mutually intelligible dialects.

The current working solution for oral proficiency testing in Arabic is for the tester to conduct the interview in MSA and to accept responses in MSA or in any colloquial dialect. Only at the Superior level is the examinee expected to demonstrate sustained proficiency in MSA.

Allen admits that the solution of writing a preliminary set of guidelines, including those for speaking and listening, based on a single variety—MSA—does not reflect the natural use of Arabic, except under special circumstances; nevertheless, MSA, potentially at least, is a means of communication between any two educated Arabs. In addition, this solution keeps Arabic in conformity with other languages for which guidelines have been developed thus far.

An alternative solution, according to Allen, would be to write guidelines for speaking and listening on the basis of the colloquial dialects, and those for reading and writing on the basis of the standard written language. This solution entails choosing specific dialect(s) and implementing Arabic courses that would offer the combination of a colloquial dialect with the standard written dialect, a practice not currently in effect in most U.S. universities, though the accepted solution at the Foreign Service Institute.

The Case of Chinese. According to Walton (1987), one of the problems with adapting the guidelines to languages such as Chinese and Japanese for use in academic testing lies in the area of reading/writing rather than in speaking/listening. The fact is that, in these languages, more time is needed to reach a comparable level of proficiency in reading and writing than in languages such as Spanish and French. This largely results from the nature of the writing system. Walton points out that exposure to a writing system, such as Chinese, does not *per se* constitute meaningful input. For instance, people who have spent some time in

France or Spain will, in all probability, learn how to write their names and addresses and to read street signs and menu items in French or Spanish without specific training. In Chinese, however, extensive training is needed to be able to perform these simple written functions. Hence the achievement of even the Novice level in reading and writing depends on a fairly protracted period of instruction and a series of developmental steps that are quite different from those involved in achieving the same level of proficiency in languages with alphabetic writing systems. Even with the refinement of the lower end of the continuum to include three subranges of Novice level, students take a long time to achieve any measurable proficiency and to move from one subrange to another.

Thus, in order to put students on the scale, the Chinese guidelines committee[4] decided to make some compromises. As a result, the Chinese reading guidelines make a distinction at the Novice and Intermediate levels between reading specially prepared materials and puzzling out authentic texts (i.e., between fluent reading and going through a sentence word by word, figuring out both word meanings and sentence structure using a dictionary).

Another concern, according to Walton, is that the testing situation constrains the elicitation of certain sociolinguistic behaviors in languages such as Chinese and Japanese. These cultures require certain language behaviors that are vastly different from those common to most West European languages. An example of such a behavior is the Chinese unwillingness to give a direct "no," preferring evasive answers that could be interpreted as "no." Concomitant knowledge of when and when not to press for definite answers is also part of such behavior. Thus, there is a need to define these features and to design situations in which they might be elicited.

Another problem Walton describes is that it may not be appropriate for a foreigner to speak to a Chinese the way Chinese speak to each other, because they

themselves expect certain behaviors of foreigners. This, of course, is a problem for all languages, but perhaps especially for languages whose cultures differ greatly from European ones. How and whether to include such expected behaviors into proficiency guidelines is an unresolved question.

The Case of Japanese. As is the case with Chinese, the Japanese Guidelines (as of spring 1987) reflect the difficulties of adapting the reading/writing guidelines to a language with a complex, nonalphabetic writing system that includes both syllabaries and characters. As in the Chinese Guidelines, descriptions of reading ability in Japanese at the Intermediate and Advanced levels distinguish between reading of specially prepared (quasi-authentic) materials without a dictionary and reading of authentic materials with a dictionary.

With regard to Japanese Writing Guidelines, a developmental hierarchy was posited starting with the ability to make limited use of the two syllabaries (Hirangana or Katakana as appropriate) at the Novice level to an emergent ability to write some Kanji (characters) at the Intermediate level.

The Case of African Languages. The problems confronting proficiency testing in African languages stem from the fact that Africa is a region of considerable linguistic diversity, representing 1,000 to 1,500 languages, and from the fact that resources for studying and teaching African languages are quite limited (Dwyer & Hiple, 1987).

The African language teaching community in the United States has initiated a number of important projects designed to coordinate the instruction efforts for African languages in this country. Two of these projects were designed to explore the application of the ACTFL proficiency model to African languages: a workshop at Stanford in 1986 and a meeting at Madison in 1987. In addition, the development of a language proficiency profiling model was undertaken.

The ACTFL workshop hosted by Stanford provided five intensive days of predominantly English-based training topped off with sessions using Hausa and Swahili. With an initial goal of examining the suitability of the ACTFL model for African languages, the workshop participants concluded that the model was based on sound principles and could provide a reliable and valid means for evaluating learners' proficiency in African languages.

In May 1987, representatives from the ten Title VI African Studies Centers assembled at a meeting hosted by Michigan State University and the University of W sconsin to rank common goals and to discuss the coordination of activities among the centers. A principal agenda item was team interviewing in much less commonly taught languages. There was near consensus on the potential for a team testing model, including a certified tester who is not necessarily proficient in the target language and a native speaker who is not necessarily a certified tester. The group suggested a three-year plan to reach this goal: (a) in 1988, a standard ACTFL workshop would be held, possibly using English, French, and Arabic as the languages of certification; (b) in 1989, a workshop would explore the design of the team approach and develop instructions for the native speaker and his or her role in the interview; and (c) in 1989, two workshops would be held using the ACTFL team approach and Hausa and Swahili as the focal languages.

A related development in the field of African languages involving the ACTFL proficiency guidelines is the "profiling" model of Bennett, Biersteker, and Dihoff (1987). This model, which grew out of a questionnaire for the evaluation of existing Swahili textbooks, was designed, according to its authors, to supplement the ACTFL global rating with a more detailed analysis of the candidate's performance for diagnostic purposes.

Problems with Tester Training

The original testing kit workshops held in 1979-80 at the Foreign Service Institute with support from the Department of Education brought interested academics together with experienced government testers in a common effort to test the hypothesis that the proficiency guidelines developed, modified, and validated over a 30-year period within the federal government had applicability in a traditional academic setting.

Seven years later, after about 50 workshops training more than 1,500 individuals in 10 languages, the answer should be clear. This does not mean that the generic guidelines are not subject to further modification as new languages representing still differing typologies are added, nor does it mean that all problems in developing compatible guidelines for additional languages and training individuals in their application in testing situations have been solved.

Less Commonly Taught Languages

It was not until 1984 that tester training became available in the less commonly taught languages. These included Arabic, Chinese, Japanese, Portuguese, and Russian. The inclusion of a particular language in this list was, of course, arbitrary, and was simply a product of whether a U.S. government tester/trainer was available to conduct initial training and whether there was interest on the part of the academic community. In the case of these languages, government trainers were available to conduct a number of initial workshops.

As a result, adequate numbers of testers in these languages were trained with new testers being added constantly to the list. In addition, through a series of tester trainer workshops conducted since 1983, ACTFL was able to terminate its dependence on government

trainers and to develop its own cadre of academic trainers not only in the commonly taught languages, but also in Russian, Chinese, Japanese, and Arabic.

Much Less Commonly Taught Languages

Recent demands from the academic sector for training in some of the nearly 160 less commonly taught languages currently available at U.S. institutions of higher education boasting National Resource Centers in Foreign Languages and Area Studies are a serious strain on national training capacity. For most of these languages, no trainer is available. This results in serious problems not only for the academic community but also for the U.S. government in general and for the Department of Education in particular. New federal legislation and companion regulations were published for public comment on Oct. 2, 1987, mandating proficiency testing for these languages. This has immediate implications for tester training as well as for adaptation of proficiency guidelines to these languages.

Using FSI trainers, ETS provided early training of Peace Corps oral proficiency testers in English for all of the requisite languages. Unfortunately, no follow-up was conducted to evaluate the effectiveness of such training. More recently, a number of academics have been trained in an intermediary language (e.g., English, French, German) while they are preparing to test in the target language (e.g., Hindi, Indonesian, Hausa, Hebrew, Polish, and Thai).

Research Needs

While the ideal training situation is target-language specific, cross-language training makes it possible to

reach languages that are otherwise inaccessible. Such training will not only extend proficiency testing to all of the much less commonly taught languages, but it will perforce provide a significant research opportunity.

Questions of interrater reliability are normally studied within specific languages. Yet, if the generic guidelines are truly generic and the training procedures are truly standardized, both rating and elicitation should be standard across different languages. The opportunity that presents itself is to study interlanguage reliability in a way heretofore not possible. A properly designed research project could provide, for the first time, empirical evidence on both the reliability of ratings across languages by the same individuals and the degree of generality among the various language-specific guidelines.

Another valuable procedure for testing in many less commonly taught languages has been used by U.S. government testers for some time. This procedure involves the presence of a tester certified in one or more languages (preferably related to the target language) and an educated native speaker/informant of that language. With appropriate direction and experience in observation during the testing of a candidate, reliable ratings can be assigned by the testing team. As in the case of cross-language training, joint testing also will provide important research opportunities.

A research agenda might include the following:

1. *A study of interrater reliability between government and ACTFL-certified oral proficiency testers.* The two groups of testers receive somewhat different training, use different guidelines, conduct and evaluate tests differently (i.e., ILR testers work in teams of two, with the final rating representing an agreement of the two testers, or the lower of the two ratings if an agreement cannot be reached; ACTFL testers conduct the interview test alone and later score it from a tape-recording). How do these differences affect ratings? For example, do ratings based on audiotape playback result in lower

ratings? (For a more detailed discussion of this question, see Clark & Lett, this volume).

2. *Intrarater reliability: An examination of differences in testing one's own students as opposed to testing someone else's.* This question is related to the larger question of testing familiar people versus strangers. In testing his or her own students, the tester usually knows their level beforehand, and at the lower levels has practiced most of the questions and answers with the examinee in class. Is it more difficult for the interviewer to show interest in the information being communicated and to interact with the examinee in a "nonteacher" manner under these conditions? Is the examinee more comfortable with an interviewer he or she knows and with whose speech he or she is familiar? How does this influence the examinee's performance?

3. *Interrater reliability: An investigation of possible differences between native and nonnative interviewers with regard to both elicitation and rating.* Are native and nonnative interviewers likely to ask different questions or to phrase them differently? Is there a difference, for example, in the way native and nonnative interviewers react to the sociolinguistic aspects of the examinee's performance? Do native and nonnative interviewers have a different "style" in conducting interviews?

4. *Interrater reliability: Differences in reliability of ratings at different levels of proficiency.* In academic situations, testers are more likely to have considerably more practice in conducting interviews at the Novice and Intermediate levels than at the Advanced and Superior levels. Does this mean that such testers would be more proficient in conducting and more reliable in rating lower-level interviews? A similar question is raised by Clark and Lett (this volume).

5. *Maintenance of rating reliability over time, particularly in the less commonly taught languages.* While maintenance of rating reliability applies to all languages, this problem may be more serious for testers in the less commonly taught languages, who are likely to

mandates competency-based language training and testing, strengthened by federal regulations affecting 93 National Resource Centers in Foreign Language and Area Studies at 53 premier institutions of higher education involving about 160 less commonly taught languages, presents monumental challenges.

As universities seek to come into compliance, there will be intense competition for the limited training resources currently available. The Department of Education, academia, and the major relevant professional associations, especially ACTFL, will need to cooperate to set realistic priorities and develop the necessary guidelines.

Which languages will be designated for priority development, and who will receive initial tester training? Will the tester training be language specific or through English or another language? Will interim procedures need to be developed to satisfy federal regulations until a sufficient cadre of testing specialists is available? Will training be extended to the precollegiate level for selected languages, and will teachers at that level receive training? Where will the required research on proficiency testing be carried out, and by whom?

What will be the role of the new federally authorized language resource centers in the area of proficiency testing? Could a small number of such regionally located centers assume responsibility for testing and tester training in their regions? How would the relationships and responsibilities of the privately funded Johns Hopkins Foreign Language Resource Center and the National Resource Centers be defined and coordinated?

The answers to these and other policy questions will, of course, require a cooperative effort by the affected constituencies. It is possible now, however, to sketch a broad outline of some of the options and factors that will influence them.

It would seem reasonable to use recent enrollment data as a general guidepost in establishing priorities. Decisions made by universities to offer courses in the less commonly taught and much less commonly taught

mandates competency-based language training and testing, strengthened by federal regulations affecting 93 National Resource Centers in Foreign Language and Area Studies at 53 premier institutions of higher education involving about 160 less commonly taught languages, presents monumental challenges.

As universities seek to come into compliance, there will be intense competition for the limited training resources currently available. The Department of Education, academia, and the major relevant professional associations, especially ACTFL, will need to cooperate to set realistic priorities and develop the necessary guidelines.

Which languages will be designated for priority development, and who will receive initial tester training? Will the tester training be language specific or through English or another language? Will interim procedures need to be developed to satisfy federal regulations until a sufficient cadre of testing specialists is available? Will training be extended to the precollegiate level for selected languages, and will teachers at that level receive training? Where will the required research on proficiency testing be carried out, and by whom?

What will be the role of the new federally authorized language resource centers in the area of proficiency testing? Could a small number of such regionally located centers assume responsibility for testing and tester training in their regions? How would the relationships and responsibilities of the privately funded Johns Hopkins Foreign Language Resource Center and the National Resource Centers be defined and coordinated?

The answers to these and other policy questions will, of course, require a cooperative effort by the affected constituencies. It is possible now, however, to sketch a broad outline of some of the options and factors that will influence them.

It would seem reasonable to use recent enrollment data as a general guidepost in establishing priorities. Decisions made by universities to offer courses in the less commonly taught and much less commonly taught

languages, as well as individual decisions by students to study these languages, already represent a prioritization, albeit implicit, and a decision that a particular language has relative cultural, economic, or political value and is thus worth teaching and studying.

Because language-specific tester training currently exists in only a handful of the less commonly taught languages, it is likely that most startup training will be through English or another language known to the prospective tester (e.g., English or French for Asian and African specialists). In some cases, it will be possible to conduct training in a language that is structurally similar to the target language, such as training in Russian in order to test in other Slavic languages. There will also be situations in which testers in one language will work together with native speakers of the target language in teams. The experienced tester may know the subject language only minimally, or may know a related language, thus being able to understand it without being able to speak it, and therefore being able to work with the native speaker in a capacity similar to that of the former linguist/informant method of language instruction—guiding the informant through the interview and making decisions as to the final rating. It is also possible that semi-direct tests of oral proficiency will be developed and validated against the oral interview for those much less commonly taught languages for which maintaining a cadre of trained testers will not be possible.

In reauthorizing Title VI of the Higher Education Act (formerly NDEA Title VI), Congress established a new Section 603, Foreign Language Resource Centers. These centers "shall serve as resources to improve the capacity to teach and learn foreign languages effectively." Activities carried out by such centers may include "the development and application of proficiency testing appropriate to an educational setting" and "the training of teachers in the administration and interpretation of proficiency tests" *(The Congressional Record,* Sept. 22, 1986).

In establishing these Foreign Language Resource Centers, Congress anticipated that giving them responsibility for direct testing and training teachers to test would present problems for individual teachers and institutions. A small number of such centers, strategically located in the United States with regional responsibilities, could significantly advance the national capacity to meet the new federal requirements for proficiency testing and competency-based training. It is also through such centers that the more limited needs at the precollegiate level can be adequately met.

In discussing the research needs, Byrnes (1987) notes that "some of the greatest benefits of the increasing work being undertaken in academia with oral proficiency testing may well lie beyond the areas that come to mind most readily, such as placement, syllabus scope and sequence, course and program evaluation, entry and exit requirements, and required proficiency levels of TAs or teachers." Rather, she sees as the most exciting prospect of the proficiency movement "its potential for giving language practitioners a framework within which to observe and evaluate the development of second-language proficiency in their students" (p. 113).

A view of the oral proficiency interview not only as a test but also as data that could yield important insights into second-language (L2) acquisition processes poses the need for language specialists to conduct second-language acquisition research, both theoretical and classroom-oriented. This research should proceed along several different lines, starting with a determination of the ultimate level of proficiency attainable under a given set of conditions (Lowe, 1985; Natelson & Allen, n.d.; Pica, 1983; Swain, 1985). Other important areas are learner variables (Beebe, 1983; Bialystok, 1983); input variables (Chaudron, 1983; Seliger, 1983); the relationship between L2 acquisition and L2 instruction (Lightbown, 1983); and the effects of formal as opposed to informal exposure on different aspects of language performance (i.e., grammar, vocabulary, fluency, and sociolinguistic and pragmatic features).

To perform such much-needed research, the prospective researchers require a background in second-language acquisition, research design, and statistics. Such a multidisciplinary background is not obtainable in highly compartmentalized foreign language departments, which typically emphasize literature and linguistics. If second-language acquisition research is to extend from ESL into commonly and especially into uncommonly taught languages, the training of language specialists must extend beyond its current boundaries of literature and linguistics to include the disciplines just mentioned. An alternative solution for performing language acquisition research in languages for which researchers with such a background do not exist is to cooperate with other departments to form interdisciplinary research teams that, in addition to a specialist in an uncommonly taught language, would include psycholinguists, educational psychologists, statisticians, and psychometricians.

Conclusion

This chapter attempts to place the development and application of proficiency guidelines to the less commonly taught languages in broader perspective. The authors have sought to highlight the significance of developing and defining the generic guidelines from their initial application to commonly taught West European languages to accommodating an increasing number of languages with widely varying typologies as less commonly and much less commonly taught languages are brought within the scope of the proficiency movement. The case is made to support the notion that guidelines are precisely guidelines, that they are dynamic and subject to modification as experience with new languages representing other linguistic typologies is accumulated. This experience may in turn have a

backlash effect on the more commonly taught languages.

Incipient experience with the much less commonly taught languages reveals serious problems in training testers and suggests imaginative and productive alternatives that hold promise for research as well.

Recent legislative action has mandated competency-based language programs and proficiency testing for the commonly as well as for the less commonly taught languages. Language resource centers will bear special responsibilities in this area and will spearhead cooperative planning and policy development in the future.

NOTES

1. The Russian Guidelines Committee was composed of Thomas Beyer (Middlebury College), Dan Davidson (Bryn Mawr College), Irene Thompson, Chair (George Washington University), Gerald Erwin (Ohio State University), and Don Jarvis (Brigham Young University). All members of the committee had either received tester training or had attended familiarization workshops. Two members (Erwin and Thompson) had previous experience as government testers.

2. The Revision Committee had two members, Thomas Beyer and Irene Thompson. Both had been active testers in the interim.

3. Vijay Gambhir (University of Pennsylvania) is a certified tester and tester trainer. Two other Hindi teachers (one at Columbia University and the other at the University of Washington) are completing their training as oral proficiency testers.

4. Members of the Chinese Guidelines Committee were Albert E. Dien, Stanford University; Ying-che Li, University of Hawaii; Chun Tan-choi, Government Language School; Shou-hsin Teng, University of Massachusetts; A. Ronald Walton, University of Maryland; and Huei-ling Worthy, Government Language School.

References

Allen, R.M.A. (1985). Arabic proficiency guidelines. *Al-ᶜArabiyya, 18*(1-2), 45-70.

Allen, R.M.A. (1987). The Arabic guidelines: Where now? In C.W. Stansfield & C. Harman (Eds.), *ACTFL proficiency guidelines for the less commonly taught languages* (Dept. of Education Program No. G008540634). Washington, DC: Center for Applied Linguistics; and Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

American Asssociation for the Advancement of Slavic Studies. (1983). *Russian language study in the United States. Recommendations of the National Committee on Russian Language Study.* Stanford, CA: Author.

American Council on the Teaching of Foreign Languages. (1982). *ACTFL provisional proficiency guidelines.* Hastings-on-Hudson, NY: Author.

American Council on the Teaching of Foreign Languages. (1984). Foreign language enrollments in public secondary schools, Fall 1982. *Foreign Language Annals, 17,* 611-623.

American Council on the Teaching of Foreign Languages. (1986a). *ACTFL Chinese proficiency guidelines.* Hastings-on-Hudson, NY: Author.

American Council on the Teaching of Foreign Languages. (1986b). *ACTFL proficiency guidelines.* Hastings-on-Hudson, NY: Author.

American Council on the Teaching of Foreign Languages. (1986c). *ACTFL Russian proficiency guidelines.* Hastings-on-Hudson, NY: Author.

American Council on the Teaching of Foreign Languages. (1987). *ACTFL Japanese proficiency guidelines.* Hastings-on-Hudson, NY: Author.

Bennett, P., Biersteker, A., & Dihoff, I. (1987). Proficiency profiling guidelines: Generic, Swahili, Hausa. Revised edition. Unpublished manuscript.

Beebe, L.M. (1983). Risk-taking and the language learner. In H.W. Seliger & M.H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 39-66). Rowley, MA: Newbury House.

Bialystok, E. (1983). Inferencing: Testing the "hypothesis-testing" hypothesis. In H.W. Seliger & M.H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 104-124). Rowley, MA: Newbury House.

Brod, R.I., & Devens, M.S. (1985). Foreign language enrollments in U.S. institutions of higher education—Fall 1983. *ADFL Bulletin, 16,* 57-63.

Byrnes, H. (1987). Second language acquisition: Insights from a proficiency orientation. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations and concepts* (ACTFL Foreign Language Education Series, pp. 107-131). Lincolnwood, IL: National Textbook Co.

Carroll, J.B. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals, 1,* 131-151.

Chaudron, C. (1983). Foreigner talk in the classroom—An aid to learning? In H.W. Seliger & M.H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 127-145). Rowley, MA: Newbury House.

Clark, J.L.D. (1986). Development of a tape-mediated, ACTFL/ILR scale-based test of Chinese speaking proficiency. In C.W. Stansfield (Ed.), *Technology and language testing.* Washington, DC: Teachers of English to Speakers of Other Languages.

Dwyer, D., & Hiple, D. (1987) African language teaching and ACTFL team testing. In C.W. Stansfield & C. Harman (Eds.), *ACTFL proficiency guidelines for the less commonly taught languages* (Dept. of Education Program No. G008540634). Washington, DC: Center for Applied Linguistics; and Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Gambhir, V. (1987). Some preliminary thoughts about proficiency guidelines in Hindi. In C.W. Stansfield & C. Harman (Eds.), *ACTFL proficiency guidelines for the less commonly taught languages* (Dept. of Education Program No. G008540634). Washington, DC: Center for Applied Linguistics; and Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Hiple, D. (1987, March). *The extension of language proficiency guidelines and oral proficiency testing to the less commonly taught languages.* Paper presented at the Indiana Symposium on Foreign Language Evaluation, Bloomington, IN.

Interagency Language Roundtable. (1985). *ILR language skill level descriptions.* Arlington, VA: Author.

Lightbown, P.M. (1983). Exploring relationships between developmental and instructional sequences in L2 acquisition. In H.W. Seliger & M.H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 217-245). Rowley, MA: Newbury House.

Liskin-Gasparro, J.E. (1984). The ACTFL proficiency guidelines: A historical perspective. In T.V. Higgs (Ed.), *Teaching for proficiency, the organizing principle* (ACTFL Foreign Language Education Series, pp. 11-42). Lincolnwood, IL: National Textbook Co.

Lowe, P., Jr. (1985). The ILR proficiency scale as a synthesizing research principle: The view from the mountain. In C.J. James (Ed.), *Foreign language proficiency in the classroom and beyond* (ACTFL Foreign Language Education Series, pp. 9-54). Lincolnwood, IL: National Textbook Co.

McCarus, E. (1987, March). *The application of ACTFL proficiency guidelines to Arabic.* Paper presented at the Indiana Symposium on Foreign Language Evaluation, Bloomington, IN.

Natelson, E.R., & Allen, D.A. (n.d.). *Prediction of success in French, Spanish, and Russian foreign language learning—An analysis of FY67-74 student data.* Monterey, CA: Defense Language Institute.

Pica, T. (1983). Adult acquisition of English as a second language under different conditions of exposure. *Language Learning, 33,* 465-497.

Seliger, H.W. (1983). Learner interaction in the classroom and its effects on language acquisition. In H.W. Seliger & M.H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 246-267). Rowley, MA: Newbury House.

Staff. (1986, September 22). *Congressional Record.* Washington, DC: U.S. Government Printing Office.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In M. Gass & C.G.

Madden (Eds.), *Input and second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.

Thompson, I. (1987, March). *Adapting ACTFL proficiency guidelines to Russian: Problems, applications, and implications.* Paper presented at the Indiana Symposium on Foreign Language Evaluation, Bloomington, IN.

Walton, R. (1987, March). *The application of the ACTFL proficiency guidelines to Chinese and Japanese.* Paper presented at the Indiana Symposium on Foreign Language Evaluation, Bloomington, IN.

Wolff, J.U. (1987, March). *The application of the ILR-ACTFL test and guidelines to Indonesian.* Paper presented at the Indiana Symposium on Foreign Language Evaluation, Bloomington, IN.

IV
Reading Proficiency Assessment:

Section 1: A Framework for Discussion

by Jim Child, National Security Agency

The assessment of reading proficiency parallels the assessment of speaking proficiency by categorizing a reader's functional abilities according to AEI reading definitions, reflecting the tasks, content, and accuracy requirements for language at each level. Performance is rated holistically by comparing it to the level descriptions.

Unlike speaking, however, reading is a receptive skill. As such, it is an ability to which testers have indirect access only. Testing must always be achieved through another channel: through speaking in a reading interview; through writing in answers to content questions; and through responses to multiple-choice items on the content of the text.

Compared with the customary classroom progress test, reading proficiency tests are usually longer in duration to permit the examinee to prove consistent and sustained ability. Reading proficiency tests also differ from classroom progress tests in the levels of difficulty covered and in the range of tasks included, since difficulty and range are both required to establish a ceiling, or upper limit of proficiency, as well as a floor. That an

Interagency Language Roundtable (ILR) test can furnish an accurate assessment of a foreign language user's reading skills has been accepted within the government for a number of years. But the fact that the test works does not explain how or why it works. This chapter addresses a number of questions related to implementing such a test.

The construct of reading proficiency is surrounded by controversy. The skill has not been fully defined, and available descriptions of the reading process vary significantly. Also, the role of texts in proficiency assessment is poorly understood and is the subject of disagreement. For example, there is a debate over the degree to which background knowledge is a factor in the student's attempt to internalize second language texts at the various levels. The theoretical underpinnings of reading assessment must be reviewed and expanded. In the next section, Phillips examines how language proficiency is viewed in the academic setting and its backlash effect on reading in the foreign language classroom. Both this and the next section reflect the state of flux of reading proficiency assessment.

This section focuses on reasons for the current debate on use of the ACTFL/ETS/ILR (abbreviated AEI) scale to assess reading proficiency. An examination of language proficiency and the language learner leads, in turn, to a discussion of the classification of texts. The section concludes with a discussion of test items and the evaluation of performance.

Teaching for "second language proficiency" is gathering increasing momentum as an instructional approach, even though the term has eluded precise definition. To judge from the heated (and occasionally rancorous) exchanges on the subject at conferences and symposia, it would seem that a number of disparate phenomena have come together in search of a consensus at any price. One group has taken its cue from the time-honored ILR descriptions of four skills and five levels, legitimized by decades of use and refined as

necessary. Another sees proficiency in quite different terms, feeling with some justification that the ILR skill level statements, as well as the ACTFL versions modeled on these, are too experiential and inadequately supported by theory. The AEI skill level definitions, argues the first group, are oriented too heavily toward classroom language study, especially adult study, as opposed to language acquisition in a natural environment.[1] But many critics of the ILR standards feel that it is premature, if not actually impossible, to impose these or similar standards on the academic community without a theoretical base for them.

Continuing research on the cognitive strategies involved in study and acquisition is to be enthusiastically encouraged. Nonetheless, from an operational point of view, the time-honored skill levels and the instruments for measuring them (especially the oral proficiency interview for speaking, and more recently, new kinds of reading tests) are effective. Moreover, the government has a persistent need to carry out wide-scale testing of employees and their families and other individuals whose language proficiency must be a matter of record.

This section refers to ILR and derivative skill level statements that have historically reflected how the language community (originally, government-agency language schools; later, academic institutions) has viewed proficiency. Because it is difficult to discuss proficiency in any but a relative way in the absence of objective criteria, a "content" paradigm is offered that is somewhat more elaborate than a text typology presented earlier (Child, 1987). Finally, the question of evaluating reading proficiency using contextual tasks is addressed. These tasks require the examinee to identify and supply missing "parts" of passages.

The learner first attempts to communicate or understand all language material at whatever level the situation requires, and does so to the extent possible relative to a text typology that is sensitized to the needs of the

target language. Both the product of the learner as he or she tries to meet language demands and the norms governing that behavior (i.e., texts graded according to type) are conceived of globally, in terms of final results. On the other hand, tasks are process-oriented: the examinee is evaluated for his or her semantic and syntactic skills while applying general language knowledge to specific textual problems.

While exhaustive descriptions of examinee behavior are available, and some materials on the gradation of texts, little exists on approaches to developing textual items (i.e., adequate analyses of item types, whether driven by form or content; the acceptability of weighting the different types; "pass-fail cutoffs").

The Language Learner

Considerable literature addresses the subject of learning a second language, both from the perspective of the student as he or she tries to develop strategies for making progress in that language, and from the experience of raters and drafters of skill level statements. Most of the literature explores the process of learning and teaching a second language. Yet the skill level statements—both ILR and ACTFL—describe in detail what learners do or appear to do once they reach each level: that is, they address the competencies attained.

As noted earlier, continuing research in the language-learning field is a prerequisite to any quantum leap in the teaching of reading a second language. Unfortunately, with few exceptions, students do not so much learn to read as to decipher. Decipherment is essentially an activity in which form triumphs over content. Exercises may train the student who has insufficiently mastered the phonomorphological system (the system of language forms as expressed in alphabets,

syllabaries, or characters) to read "word for word" to the detriment of meaning, thus causing the means to interfere with the ends. Other kinds of drills may support decipherment at the expense of reading in a different way; students may be asked to parse sentences with the aim of attaching formal labels to words conveying messages. Skillfully planned curricula that place content before form would go a long way toward changing the relatively entrenched discrete-item approach (Phillips, 1984).

A change of this sort would in addition comport with the natural human tendency to generate or process texts. No such natural tendency to perform grammatical analysis exists. The entire constituency of language users (native speakers and learners), psychologically prepared to speak or read for meaning, has little motivation for grammar exercises. The factors limiting proficiency from this perspective are degree of verbal intelligence, technical and/or cultural background knowledge, and ability to use various strategies, such as circumlocution, in the absence of adequate control of the second language.

While general intelligence is not open to intervention, background knowledge in the cultural and technical domains can certainly be improved. Each learner brings to the target language a particular fund of knowledge about the world that contributes to his or her progress. What is often not considered, however, is the extent to which various kinds of knowledge affect language behavior at different levels in second language learning. To the extent that it has been an issue at all, subject-matter control has been a source of concern with respect to potential effects on test results. An examinee may appear (especially in an oral proficiency interview) to know the second language better than he or she actually does. This phenomenon is known to linguists as "semantic feedback" and to testers as the "hothouse special." Knowledge of subject matter, rather than command of the target language system,

determines the performance of the examinee and interferes with the evaluation of proficiency in the testing situation. Yet in the realm of practical applications, content-oriented skills may be a *sine qua non.* One of the most useful approaches to deciphering unfamiliar material is to gain currency on the topic or topics to be processed by studying these in one's native language first, and then to transfer this content knowledge to reading the material in a second language.

Another strategy for the learner who does not control a particular construction or lexical item is to use circumlocution, which ensures communication at the expense of precision or elegance. While an examinee may be marked down for inaccuracy, a person using language "in the real world" may salvage a precarious situation, fully justifying the mastery of this strategy.

A great deal of research is needed on the effects of technical and cultural knowledge and circumlocution skills on performance. It will be useful to explore ways to measure these skills' contribution to communication and to enhance them, given the virtual certainty that they will be needed in the future.

Classification of Texts

To establish the reading skill level of a language learner, a theoretical construct is needed that is applicable to all written languages and provides an objective, common yardstick of evaluation. A criterion of this sort has always been implicit in the ILR statements, even though it is not fully appreciated. The question is dealt with at length by Child (1987), but two examples are noteworthy here.

The lead sentence of the ILR statement on reading for Level 2 is worded as follows:

> Sufficient comprehension to read simple, authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context.

The main focus of this (and every other) ILR reading statement is the mental activity of the reader, without speciying what "simple authentic written material" might be like in one language or another. Although the description is further qualified as "straightforward, familiar, factual material," what is considered straightforward and factual may differ from one language to another. What is factual or straightforward to readers in one culture may appear to be propaganda to those of another. However, philosophical differences as to what constitutes factuality aside, the direct presentation of material that is amenable to reportorial or narrative treatment (regardless of its truth value) falls under the "instructive" rubric of textual Level 2.

From the formal standpoint, the question is the degree of complexity of expression in which the content is embedded (i.e., the syntax). Generally, the syntactic patterns at Level 2 are high-frequency structures with minimal information, in keeping with the semantic and pragmatic expectations of the level. Element order within sentences is rarely atypical. For example, if the normal order of the declarative sentence in a language is subject-object-verb with variations permitted only for special rhetorical purposes, the S-O-V arrangement will appear almost exclusively in Level 2 texts because these (at least in written forms) rarely involve affective or evaluative material. As for morphology—the system of affixes or "function words"—these elements are usually obligatory and are critical for accuracy (even when expendable in lower-level communication) both within sentence or clause and at the discourse level. Finally, intonational patterns, clearly marked in speech but underrepresented in writing, are also obligatorily supplied for phrases in sentence-long units.

A second example is from reading Level 4. According to the ILR statement, the person reading at this level is "able to read fluently and accurately all styles and forms of the language pertinent to professional needs." This description, admittedly vague, assumes an extensive command of the target culture (both in its narrower and broader senses) as well as whatever content domains are controlled by the examinee who "is able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references." Depending on the topic, the language may be highly metaphorical or allusive, with elaborate rhetoric or a deliberately crafted lexicon meant to convey worlds beyond the words. The Child (1987) typology subsumes such varied products under the Projective Mode as follows: "Shared information and assumptions are at a minimum and personal input is paramount."

Formally, virtually any syntactic device, but especially those of low frequency and high information content, may be encountered, again according to the supposed needs of both writer and reader. For instance, low-frequency word order (perhaps combined with an infrequent lexical item) may be used to achieve a special effect (e.g., "His objections notwithstanding" for "Despite his objections"). By definition, obligatory morphemes are as important at Level 4 as elsewhere, but the cohesive devices (reference, substitution, etc.) that are immediately clear in Level 2 texts often require interpretation at the higher level. Synonymy, too, makes much greater demands vis-à-vis lexical cohesion than at Level 2, as individual stylistic choices—the major characteristic of level 4—come into play.

Limited as these examples are, they illustrate what learners can do at two stages of attainment, especially when combined with intralanguage textual descriptions anchoring the behavioral statements. As noted earlier, both documents reflect results, encapsulating the end product of the language-learning process.

Test Items and Evaluation

Test design and evaluation, especially in a contextual format, are extraordinarily difficult tasks. First it is essential to create a sound balance of item types at each textual level tested, ensuring that these are distributed relatively evenly throughout the exercise. Then decisions need to be made on "pass-fail cutoffs," determining how many mistakes, and of what kind, are allowable at each level.

The experience of most teachers and test designers in the second and foreign language field has been dominated by formal grammar. Traditionally, parts of speech are tested in terms of case endings for nouns, tense markers for verbs, adjective-noun agreement, and so on. In more recent approaches to test design and evaluation, the content of a passage may be the primary focus of the exercise, with formal features included to establish accuracy.

A means to reconcile form and content is critical to the entire concept of proficiency testing. In an experiment at the National Security Agency, differentially weighted items are provided at Levels 1^+ and 2^+. At Level 1^+, high-frequency vocabulary items, especially verbs and nouns, were deleted, as they interact in case frames within sentences ("case" here refers to the logical relations obtaining between noun and verb as well as among nouns; such relations may or may not be formally indicated, depending on the typology of the language concerned). The aim is to elicit the examinee's knowledge of the forms of the language being studied. At Level 2^+, items involving the linkage among verbal and nominal forms are, of course, deleted. In addition, however, extended textual segments that contain formal cohesive elements are deleted: items that refer to nouns and verbs elsewhere in the text. These include mainly third-person pronouns, adverbs such as *thus* and *so,* conjunctions and even verbal forms such

as *does* and *is* and nouns such as *thing, matter,* and others. Testing such cohesive forms is a way to determine whether the examinee is truly following the argument (i.e., the content) of a given text in a second language. Because form and content are reasonably congruent in most languages up to Level 2+ (i.e., concepts are expressed in roughly predictable ways, with a limited number of variations), form and content are testable together. At higher levels, where considerable intellectual or esthetic creativity comes into play, the expression of content may run from the (deceptively) simple to the highly complex. Testing texts at these levels require a degree of sophistication that does not lend itself to thinking in terms of items.

Unfortunately, the practices of teaching and testing have resulted in a dichotomy between form and content. Thus it often happens that learners have been force-fed on paradigms and mastered them to a surprising degree, only to fail completely in processing simple texts. By the same token, some learners have acquired their second language in natural language-use environments and are content-oriented as a result. Such learners may have little trouble following the main argument of a text (except for vocabulary items that are seldom encountered in spoken language registers) but may have trouble restoring grammatical deletions, an important exercise in all language batteries.

These considerations led to the idea of developing "mixes" of items and weighting them differentially. Because the problem of different types of learners is not unique to one institution, but pervades the foreign and second language education community, the analysis of item types and their organization into contextual tests must go forward.

Summary

Professionals concerned with the assessment of reading proficiency must come to terms with the interaction between reader and text. Readers bring to their interpretation of texts various degrees of knowledge of the world and control of the grammar and lexicon of the target language. Texts, on the other hand, are the result of native speakers reporting facts, narrating events, and commenting on topics of interest. Such texts obviously represent appropriate material for evaluating problems of form and content peculiar to each language and for determining textual levels of difficulty. Once these levels are determined, appropriate tests can be designed and administered to assess examinees' proficiency.

NOTE

1. *Study* is distinct here from *acquisition; learning* is used as the generic term for gaining control of a second language regardless of the method.

References

Child, J. (1987). Language proficiency levels and the typology of texts. In H. Byrnes & N. Canale (Eds.), *Defining and developing proficiency*. Lincolnwood, IL: National Textbook Co.

Phillips, J.K. (1984). Practical implications of recent research in reading. *Foreign Language Annals, 17*(4), 285-296.

Section 2: Interpretations and Misinterpretations

by June K. Phillips,
Tennessee Foreign Language Institute

While speaking proficiency assessment gained fairly easy entrance into academia, the reception given reading proficiency assessment has been cool, to say the least. In addition to the reasons discussed by Child in the first section of this chapter, two further reasons are discussed here. First, academia has always concerned itself with reading, devised ways to teach it, and conducted considerable research into the nature of the reading process. Thus, the reading proficiency definitions, unlike the oral proficiency interviews, do not fill a void but must contend with numerous other views about testing reading and the nature of the reading process. Second, the tendency in schools and colleges to deal with higher-level texts before the requisite skills are mastered by the student leads to entirely different views of reading skills than the view suggested by the ACTFL/ETS/ILR (AEI) proficiency guidelines. The major controversy has been formulated in the question, "proficient readers or proficient texts?" This chapter argues that this opposition is counterproductive. This is but one of many significant issues in adapting reading proficiency assessment to the academic setting.

Reading proficiency is discussed here in terms of both texts and the reading process, and implications of the AEI reading definitions (including texts and process) are explored for the teaching of reading in the

foreign language classroom. After a discussion of needed research, the section ends with observations on the present situation and directions for the future.

Reading skills as described in the AEI definitions parallel those in other modalities in that they address a continuum of tasks that native readers are capable of performing. They differ in that learners must do more than manage the language they control as output; they must deal with the input provided by a writer who is not attentive to their level of language. The continuum, therefore, must consider both texts and tasks. Those who have claimed that the definitions for reading address only text type have not read them carefully or thoroughly. As Child suggested, the inclusion of statements about reader behaviors and learner strategies contributes to the complexity of the level descriptions and current procedures for assessment. Thus, the descriptions address proficient readers and not just proficient texts, contrary to Bernhardt's (1986) interpretation. Such mixing of diverse elements in the AEI definitions calls for refinement. As the profession grows in its knowledge of receptive processes, confirms or rejects models of second-language reading, and develops more sensitive measures of comprehension, the definitions should be modified to reflect these gains.

The Texts to Be Read

In their present form, the reading proficiency definitions can enhance traditional reading experiences in several ways; primary among them is the delineation of common text types according to the impact they have on readers. At one time, in the academic setting, it was common for most foreign language learners to deal with a narrow range of texts, usually 200-250 words long, in a recombination or highly controlled narrative. The small number of students who became language

"majors" would enter the world of literature, ideally to understand and appreciate, but more realistically to decode and manipulate. Most important, neither group ever attained a level of proficiency that encompassed more than a few strata of the texts available to the native reader or tasks resembling what the second language reader can do in his or her first language. Academia must now make a conscious decision to develop a curriculum that includes a wider range of reading skills and passage types. The alternative is to maintain a selective focus in which readers may become proficient with limited tasks in specific texts. Reading proficiency measures are not appropriate or useful in the latter case, because the hierarchical assumptions of the AEI definitions cannot be met. The reading comprehension tasks in which passages meet the criterion of a single type (see Child s [1987] typology) produce a "score," whereas a reading proficiency test according to the AEI scale demonstrates the reader's array of skills. This is because when a sustained level is assigned, the examiner assumes that lower-level functions and texts can be handled successfully.

It is important to note that the hierarchy suggested by the reading definitions differs in reality from that of speaking, from which it was derived Acceptance of the definitions implies the intention to include the spectrum of skills in instruction and evaluation. In reading, the hierarchy is not confirmed automatically by student performance, whereas in speaking, no learner can "narrate, describe, or speak in paragraphs" (Avanced/2) who cannot "maintain simple face-to-face conversations" (Intermediate/1). Independent of method or materials, learners measured over time perform in speaking at Novice/0 levels with words and learned materials before they demonstrate creativity in sentence-length structures (Intermediate/1). Schooling can, on the other hand, produce Advanced/2 readers who may well lack the visual literacy of the Novice/0, especially if their skills have been developed in a language-centered program. Reading is not a natural

skill; it is learned. Thus the presence or absence of the hierarchical assumptions for reading depend on the goals of the curriculum.

While it may be possible to teach students to read only segments of the continuum mastered by native readers, such comprehension should be defined solely in terms of passage types read. Otherwise, students presume that their ability to read an edited text and score highly on a reading comprehension test means that they can "read." To their amazement, their first experience in the target culture often vividly demonstrates that reading the realia around them is not as easy as they imagined. The task of defining or assessing reading proficiency within the AEI paradigm imposes the necessity to demonstrate that the hierarchy applies to reading outside the classroom, and that any rating implies successful reading of lower-level texts. Thus, test passages must be representative at all levels.

Processes and Strategies for Reading

The continuum of text types on the reading scales is only part of the description contained in the AEI definitions. Just as for the other skills (see Lowe, this volume, Note 4), functional trisections have been charted for reading to clarify dimensions embedded in the level descriptions. For reading, one trisection includes text type, reader function, and reader strategies (Canale, Child, Jones, Liskin-Gasparro, & Lowe, 1984). Two additional segments that may be considered are author intent and author accuracy.

This breakout is important in identifying strategies readers use and the functions they demonstrate with the various texts. On its surface, the listing of strategies horizontally with text types could be confusing if interpreted to mean that these strategies are uniquely associated with the level at which they appear. In reality,

effective readers use all these strategies at all levels in their efforts to assign meaning to a text. Actually, the side-by-side placement in the trisection, as well as the mention of specific strategies in the narrative descriptions, indicates that consistent control of that strategy correlates well with the accomplishment of the function required. For example, with "orientational" Intermediate/1 level texts, the reader usually needs only skimming and scanning strategies to identify the main ideas. This strategy tends to be the most important for Level 1 functioning.

In describing the reading process in general, these same labels are used to designate learning strategies. Consequently, an individual reader aiming for comprehension of an orientational text (e.g., names of stores, street signs, travel forms) may also use inferencing or contextual guessing strategies to assign actual meaning.

To summarize, the AEI definitions, through the rather tightly packed descriptions, provide insights into the tasks and accuracy of comprehension that readers demonstrate as they interact with a range of texts. For teachers, this knowledge is a basis for choosing materials, particularly authentic ones, and for developing activities that set reasonable tasks given the reading levels targeted for their students.

Unlike oral proficiency testing, for which no suitable testing mechanism existed, academia has always tested reading in some form using a wide variety of methods. As a result, the task of introducing tests of reading proficiency according to the AEI scales presents problems, because traditional tests have largely required decoding parts of passages, rather than demonstrating consistent and sustained control of ILR functions and of the content specified by the AEI definitions. In fact, a mismatch exists between what the AEI scales define as reading and as suitable reading content and what academia has traditionally regarded as reading and suitable reading test material. The problem is not that AEI reading proficiency tests may not furnish generally

valid and reliable results, but rather that the classroom teacher will, without an introduction to the system and how it affects curriculum, be baffled by it. If, as has been maintained, teachers tend to teach for the test, two changes must occur for reading proficiency tests to make sense in the classroom: (a) new strategies must be taught and (b) new types of material must be included. Once these changes are introduced, then teacher and student alike will understand more readily how an AEI reading proficiency score reflects the examinee's ability.

Implications for Teaching Proficiency in Reading

The decision to integrate concepts of the reading guidelines into the foreign language curriculum entails identifying materials and learner strategies that surpass the current level of performance. Several issues concerning both texts and strategies must be addressed.

A Wider Range of Tests

At the start of foreign language classes, students should develop skills in reading Novice-level materials at the lowest end of the continuum that reinforce or enrich the topics being studied. These selections are highly contextualized and often require some background knowledge on the part of the student. Lacking that, the teacher may need to provide cultural contexts, advance organizers or other activities that facilitate the student's access to the text. Many basic textbooks contain bits of realia and cultural information, but these need to be converted from decorative pieces to sources of information. The basic text must usually be

supplemented with a collection of reading materials. Most beginning courses follow a topical presentation. To find appropriate materials, the question to ask is, for example, "What does one read in French or German or Arabic that contains expressions of weather, or colors, or clothing, or food?" Appropriate texts at this level do not display the characteristics of paragraphs but are in the form of symbols, short phrases and lists. It is relatively unimportant if students cannot understand every word or detail; recognition of essential information appropriate to the task suffices and stretches the learners' thinking beyond the mechanical processing of artificial documents. The abundance of Novice/0 level texts is attributable to the heavy reliance among the world's societies on signs, symbols, advertisements, announcements, and so on.

As students ascend the scale, authentic materials representing other test types should systematically provide for "real-world" reading practice. At the Intermediate level, fewer materials are available, because this type of fairly simple, straightforward passage is of limited use to native readers. Good materials can be found in subject areas related to "survival" topics, particularly shopping, the post office, banking, transportation, food, and lodging. Brochures with good visual support, though aimed at native speakers, try to simplify both language and procedures for new customers. Another resource is general-audience magazines, but the teacher may need to specify how much and what parts of articles are at the level of challenge. Materials must not be so difficult that they cause frustration.

The large majority of texts used by native readers is at the Advanced/2 level and above. Teachers and authors of foreign language textbooks play an important role in selecting the actual texts, for these must reflect content, interest, and language that is or can be made accessible to the student. The key to accessibility lies in deciding what the learners can be asked to do and helping them to do it. It is more a matter of editing the *task* than editing the *text*.

Teaching Effective Strategies

Before a wider range of texts will lead to more proficient reading levels, the process of reading must also be taken into account. At beginning levels, when "foreign" describes the culture and context as well as the language, successful reading depends on the learner's development of a set of effective strategies. A wide spectrum of techniques have been proposed and explored (see Grellet, 1981; Omaggio, 1986; Phillips, 1984; Swaffer, 1983 for overviews). At a minimum, they involve some degree of prereading activities, skimming or scanning phases, and comprehension guides to verify that the reader can carry out a similar task in real life.

Most recent research has confirmed the importance of prereading activities for students. Acting as advance organizers that activate students' abilities to predict, anticipate, guess from context, absorb into existing schema, or simply recall background knowledge or experience, this first stage is critical to comprehension of authentic materials. (For examples of research studies, see Carrell & Eisterhold, 1983; Hudson, 1982; Haus & Levine, 1985; Steffenson, Joag-Dev, & Anderson, 1979.) Teachers who are aware of their students' linguistic, cognitive, and experiential background are in the best position to determine effective prereading activities. Furthermore, they can best judge how much preparation is required for entry into a specific passage. The teacher must walk a narrow line between providing too much and too little help.

Passages do not exist at as many levels as there are differences in student ability. The proficiency descriptions can serve as a tool for designing tasks and determining specific areas that can be tested for comprehension. Tasks requiring intensive in-class reading dominated in the past, but more recent materials include texts that are to be scanned only for main ideas, supporting detail, or designated pieces of information.

Using these procedures, many passages that were once considered too difficult become readable.

Instruction that is built on these principles, derived from the definitions, blends concern for the reading process, the learner's repertoire of effective strategies, and exposure to texts representative of those in the target language. Students exiting such programs will have a chance of improving their reading skills as they learn new information about their environment and that of others. More important, they will be able to manage reading tasks that are carried out for real work purposes; their skills will allow them to enter the spectrum of readings available to native readers so that they no longer exhibit mixed proficiency, jumping from level to level in a nonfunctional way.

While the implications of AEI reading proficiency testing for the classroom continue to be explored, tests in use, though not thoroughly understood, nevertheless seem to work. It should be reiterated that AEI reading proficiency testing has proven effective, as evidenced by its use in government for more than 30 years. Yet the ILR has devoted less attention to reading than to speaking, and little understanding exists about the interrelationship of functions, content, and accuracy in assessing reading proficiency. These facts compel more research and underscore the urgency of the need. Perhaps this research will explain why the AEI reading system scales work, especially when the content from one test passage to another varies greatly and examinees often only partially exhibit the skills set forth in the AEI definitions.

The Research Agenda for Reading

The reading descriptions and the implications that academic teachers and administrators draw from them must undergo tests of applicability, assessment,

and rigorous experimentation. Assumptions must be challenged, and the practices being advocated must be tested to determine whether they have positive effects on comprehension. The implementation of a wide-ranging research agenda cannot be delayed. Lett and Clark (this volume) address several relevant issues, and what follows is an addendum to that chapter.

First, valid tests of comprehension of passages identified by level must be developed specifically for the academic world. A program to assess whether performance at one level can guarantee that the lower levels are also controlled would provide evidence of the existence of a hierarchy for those who learn to read a foreign language. Within level descriptions, research to confirm whether the trisectional statements appropriately combine text, function, and strategy could also be conducted. Claims that the descriptions must be validated should be answered, but the face validity has been established by the historical derivation of the model; that is, they are the result of analyzing learner performance. Research is probably not capable of producing a set of guidelines because no single theory could generate them. The AEI definitions deal with specific linguistic items and how and to what extent learners comprehend them. Research can, however, confirm or reject their viability as good descriptors of performance.

Another agenda item is classroom-based research on the interactions that occur between reader and text to determine whether instructional strategies render texts accessible. Every aspect of prereading, the effects of advance organizers, or applications of schema theory (Minsky, 1982) should continue to be explored. Experimentation with newer kinds of test items or formats for comprehension would also be valuable. Computer programs that provide help menus that respond to student prompting would tailor intensive reading practice to the individual learners' needs, a major advance over the full-class problem-solving that usually occurs. The ultimate research and development effort would be to develop a computer-adaptive reading test that could

efficiently assess the range of texts a student can process to given levels of accuracy so that a rating given a student conveyed the kind of information that the oral proficiency interview does (Carton & Kaya-Carton, 1986; Dandonoli, 1987). Students would then have a fairly accurate idea of what they can read and what the next stage requires.

Interpretations, misinterpretations, implications for the classroom, the need for research—all influence how AEI reading proficiency testing will be received. This chapter has come full circle—from testing room to classroom and back. While the starting point and end point are the same, the understanding of that point changes significantly when the AEI definitions are studied with new eyes. Frankly, examiners in the past have tested reading achievement, the presence of bits and pieces of content and ability. What the AEI definitions suggest be tested in the future, particularly when students are taught for the test, is proficiency; that is, consistent and sustained ability to read a wide variety of texts with appropriate reader strategies.

References

Bernhardt, E.R. (1986). Proficient texts or proficient readers? *ADFL Bulletin, 18*(1), 25-28.

Carrell, P., & Eisterhold, J. (1983). Schema theory and ESL reading pedagogy. *TESOL Quarterly, 17,* 553.

Canale, M., Child, J., Jones, R.L., Liskin-Gasparro, J.E., & Lowe, P., Jr. (1984). The testing of reading and listening proficiency: A synthesis. *Foreign Language Annals, 17,* 389-391.

Carton, A.S., & Kaya-Carton, E. (1986). Multidimensionality of foreign language reading proficiency:

Preliminary considerations in assessment. *Foreign Language Annals, 19,* 95-102.

Child, J. (1987). Language proficiency levels and the typology of texts. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency.* Lincolnwood, IL: National Textbook Co.

Dandonoli, P. (1987). ACTFL's current research in proficiency testing. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency.* Lincolnwood, IL: National Textbook Co.

Grellet, F. (1981). *Developing reading skills: A practical guide to reading comprehension.* Cambridge, England: Cambridge University Press.

Haus, G.J., & Levine, M.G. (1985). The effect of background knowledge on the reading comprehension of second language readers. *Foreign Language Annals, 18,* 391-397.

Hudson, T. (1982). The effects of induced schemata on the "short circuit" in L2 reading: Nondecoding factors in L2 reading performance. *Language Learning, 32,* 1-31.

Minsky, M. (1982). A framework for representing knowledge. In J. Haugeland (Ed.), *Mind design.* Cambridge, MA: MIT Press.

Omaggio, A.C. (1986). *Teaching language in context: Proficiency-oriented instruction.* Boston: Heinle and Heinle.

Phillips, J.K. (1984). Practical implications of recent research in reading. *Foreign Language Annals, 17,* 285-296.

Steffenson, M.S., Joag-Dev, C., & Anderson, R.C. (1979). A cross cultural perspective on reading comprehension. *Reading Research Quarterly, 15,* 10-29.

Swaffar, J.K. (1985). Reading authentic texts in foreign language: A cognitive model. *Modern Language Journal, 69,* 15-34.

V
Issues in Writing Proficiency Assessment

Section 1: The Government Scale

by Martha Herzog,
Defense Language Institute

Writing has never been given great emphasis in government language evaluation. The original effort in the 1950s to inventory, define, and measure the language proficiency of government employees led to a system covering all four skill modalities. However, the testing that evolved concurrently with the language skill level descriptions focused on speaking and reading, with some attention paid to listening comprehension (see Sollenberger, 1978). For many years, the writing proficiency scale was primarily used only for reference.

While their curricula may require some writing, the government language schools do not test this skill as part of the end-of-course evaluation. Furthermore, any writing done during the course is not evaluated according to the proficiency scale.

In about 1980, however, the Civilian Personnel Office of the Defense Language Institute (DLI) took a careful look at writing proficiency. In the hope of improving the assessment of prospective teachers'

language ability, the personnel specialists proposed to test all four skills according to the government scale. This was a pioneering effort, in advance not only of the ACTFL guidelines project and the proficiency movement, but also of the DLI Test Division's major undertakings for measuring students' language proficiency. At first, the personnel office concentrated on training testers for English, but their ultimate goal was to create a sizable cadre of testers in all languages taught at DLI. This goal has been accomplished. DLI has more than 250 certified testers in 26 languages, serving both the personnel system and the Test Division. Though fully trained to assess writing proficiency, these testers rate this skill only for the personnel office; under normal circumstances, they are the only federal employees who actively use the writing level descriptions. The present focus is on the one practical use of the writing scale.

By the time DLI became seriously involved in proficiency testing in 1981, the Interagency Language Roundtable (ILR) Testing Committee had identified a need to revise and expand the level descriptions for all four skills. Revisions were needed for several reasons. First, there were no written descriptions of the plus levels despite their widespread use. The Foreign Service Institute had developed helpful guidelines for awarding plus ratings in speaking (Adams & Frith, 1979), and the agencies shared an informal tradition concerning the general use of plusses. It seemed time to agree on complete descriptions for all 11 points on the scale.

It was also time to provide a fuller explanation of the functional ability found at each level. This was not a matter of a need for change; rather the need was for amplification. Examples were needed for the speaking descriptions. Within the reading descriptions, a distinction had to be made between texts at a specific level and people who read at that level. Also added was an explanation of the limited functions a Level 1 or Level 2 reader can perform with higher-level texts.

Another problem that interested DLI involved commensurability among skills. For example, experience

and analysis had shown that the standards for attaining a Level 1 in reading were somewhat lower than those for speaking. As the issue of commensurability was discussed, the writing descriptions were carefully scrutinized. DLI's participation in these discussions may have marked the first analysis by an agency that actually proposed to use the writing scale for testing. It turned out that the writing descriptions were not as carefully graduated as those for other skills. Thus, while a Level 1 speaker could perform certain limited language tasks independently and the Level 3 speaker could function effectively in both social and professional situations, all writers below Level 5 required the assistance of an editor.[1] Inadequacies in the writing descriptions particularly pointed up the need to revise the system to make it commensurate.

Several joint government and academic projects were begun at a DLI testing conference late in 1981. Among them was an on-the-spot exercise by Pardee Lowe Jr. and Adrian S. Palmer to remove the most glaring deficiencies from the writing descriptions. During a two-day period, they rewrote the base-level descriptions to bring them into line with speaking. Most of this work found its way into the revised ILR level descriptions that were finally published in 1984, and into the ACTFL guidelines for writing.

This important effort by Lowe and Palmer paved the way for revision of the entire scale and should not be minimized. Nevertheless, in retrospect, it appears that the ILR Testing Committee might have devoted attention to additional improvements. During a year of intensive rewriting, it was clear that the Testing Committee assigned the lowest priority to descriptions for writing.

As noted earlier, even the improved writing scale is little used within the federal government. If the military services were to ask DLI to test students, its certified testers could certainly evaluate the writing samples; they receive good training in the use of the scale. However, DLI's Personnel Office remains the only client for such testing.

Applicants for teaching positions are tested in both English and the target language. The lowest acceptable score is a Level 2 in English (Advanced on the ACTFL scale) and Level 3 in the target language (ACTFL Superior; for a comparison, see the figure in the Introduction, p. 4, this volume). The present writing test consists of a single essay answering the question, "How has your background and experience prepared you for employment at DLI?" The essay is rarely written under controlled conditions, and it is tacitly assumed that one version—English or target language—will probably be a translation of the other. Because the topic has not varied for several years, those who choose to compromise the test are in an excellent position to do so.

The greater problem, however, is that this assignment severely limits educated, skilled writers' attempts to show versatility with prose style or their ability to develop a complex idea. A superficial glance at the scale would suggest that even the best writer could handle the topic adequately at Level 3. Careful scrutiny of the descriptions reveals an even more fundamental difficulty. Level 3 implies both formal and informal style, tone, and subject matter; clearly, no single essay can demonstrate this range of ability.

A Direction for Future Tests

Because of higher priorities for test development, DLI has not yet developed a meaningful writing test for prospective teachers. However, the time seems right to begin work.

An extremely interesting prototype test that covers the ACTFL scale from Novice through Superior has already been proposed by Magnan (1985). She designed the prototype to be administered in two sessions, each lasting about 2.5 hours. The testing time is lengthy, but no surprise to those familiar with the guidelines or those

who have administered essay exams under controlled conditions. Magnan's model succeeds in suiting topics appropriately to proficiency levels and in demanding sufficient variety of topic and task to fully meet the requirements of the ACTFL guidelines. If DLI were to test students' writing ability, it would afford an excellent chance to experiment with this model and indeed with many of the suggested topics. Magnan's test should be given a thorough trial, particularly at institutions that implement a proficiency-based curriculum.

An earlier prototype had been developed by the author at DLI covering Levels 2 through 5, the portion of the scale used by the personnel office. It is intended only for English-language testing. Tests of similar design and length should be generated in the various target languages by teams certified to test in each language. At the higher levels of proficiency, it is doubtful that topics and tasks can be designed that will work successfully in more than one language. Eve. · aspect of the test, including the instructions, must be sensitive to the demands, problems, and conventions of writing in the specific language. If this assumption is too cautious, a research project suggested later in this chapter will show this.

Several certified testers were consulted in the development of the English prototype.[2] To test through Level 5 requires samples of formal and informal writing and a "variety of prose styles pertinent to professional/educational needs." The minimum number of writing assignments and topics and the minimum length of each assignment that would enable examinees to demonstrate their true writing ability had to be determined.

It was assumed at first that one assignment, with a choice of topics, would be needed for each level tested. However, DLI's extensive experience with testing reading between 1982 and 1986 revealed that a carefully selected text with well-designed questions can measure reading comprehension skill accurately at several proficiency levels. For example, DLI has been able to use

Level 3 texts in a cloze format to discriminate precisely among readers who scored at Levels 1 thr ugh 3^+ in a thorough face-to-face test of reading. On the basis of this experience, it can be hypothesized that writing tasks aimed at Levels 2, 3, and 4 will furnish samples that allow discrimination among writers through the upper part of the scale.

Like Magnan, DLI was concerned about test length. But in addition to the problems she anticipated, DLI realized that job applicants are not as readily available as students. Therefore, a single testing session was needed. And the concern had to be to determine minimum, rather than optimum, length.

A carefully designed and validated test becomes both expensive and valuable. To preserve the investment, the test had to be administered under controlled conditions at DLI or another government office. DLI's assignment of topics had to take this factor into account. Use of reference materials such as dictionaries, thesauruses, or style manuals could not be permitted unless their availability at every testing site could be guaranteed. In addition, DLI sought to create topics that no category of applicants would be likely to have prepared for in advance. At the same time, the topics had to be of interest to the average applicant. The task of narrowing the test to measure writing proficiency according to the level descriptions and nothing more—not general knowledge, professional preparation, intelligence, or retention of published articles—was far from easy. And until the test can be validated and tried, it can by no means be assumed that all the problems have been resolved.

The proposed test appears in Figure 5.1. Part I presents a straightforward Level 2 task. This particular topic covers the reference in the level descriptions to "routine social correspondence" and would show ability to write about "daily situations." It assumes organizational and discourse ability at the paragraph level only. The instructions mention grammar and vocabulary because the level descriptions are relatively specific about

The following assignments will be used to evaluate your English writing proficiency according to the ILR language proficiency descriptions. Pay careful attention both to the topic and the intended reader.

PART ONE

- Assume that you have just returned from a trip and are writing a letter to a close friend. Describe a particularly memorable experience that occurred while you were traveling.
- This will be one paragraph in a longer letter to your friend. The paragraph should be about 100 words in length.
- You will be judged on the style and organization of this paragraph as well as vocabulary and grammar. Remember, the intended reader is a close friend.

PART TWO

- Imagine that your reader is a young American—20-25 years old, bright, educated, and outgoing. Choose one of the following topics. In about 300 words, write about the subject in a way that will be interesting to this reader.
- You will be judged on the style, organization, coherence, and complexity of your essay as well as the richness and precision of vocabulary and the accuracy of grammar and spelling.

A. Explain why it would be valuable to learn a second language.
B. Describe a person who had a great influence or your life. Explain why this person was so important to you.
C. Present your own definition of success, giving reasons.

PART THREE

Assume that you have been asked to write a paper to be presented during the annual meeting of a professional organization to which you belong; it will later be printed in their quarterly newsletter.

- Select one of the following topics and write a paper of approximately 750 words.
- You will be judged on the style, organization, logical development, and complexity of your paper as well as the richness and precision of vocabulary, accuracy of grammar and spelling, and the suitability for the intended audience.

A. Teachers' resistance to change.
B. The influence of television on language skills.
C. Quality versus equality in higher education.
D. The move toward neutralizing gender in language.

Figure 5.1. Proposed Test of Writing

the accuracy expected at Level 2.[3] However, because the examinee is urged to pay attention to style and to address a particular audience, the evaluator can also use this assignment for part of the overall rating of higher-level examinees. Experience has not shown that Level 2 writers themselves control style or tailor their writing for an assumed audience. However, reliance on the single DLI topic—discussion of past experience—may have the effect of limiting examiners' knowledge of the Level 2 writer's ability.

Part I should certainly screen out examinees whose performance falls below Level 2. However, it may not provide an adequate sample to discriminate between the true Level 2s and the Level 2+s.

Part II should do a great deal more. The choice of topics should allow examinees to write on subjects they have thought about before, without offering any likelihood of specific preparation for the test. The true Level 2 could probably write well enough to confirm the evaluation of the first assignment. However, this second assignment should establish a ceiling for the current proficiency as a single paragraph might not. The solid 2+ writer could use this assignment to demonstrate an ability to "write about concrete topics relating to particular interests and special fields of competence."

After completing Parts I and II, examinees whose proficiency level exceeds 2+ will have shown their higher-level ability through their style, organization, and tailoring of the subject to the audience. In most cases, grammatical control and a sufficient difference in style between Parts I and II would assure evaluators that an examinee is at least Level 3. Trial will be required to determine how much discrimination can be made on the basis of these two essays.

In combination with the first two assignments, Part III should provide enough information for assigning proficiency levels at the top of the scale. Any of the four topics can be considered pertinent to professional interests and needs. Admittedly, they are extremely general; however, generality should ensure lack of

conscious preparation for the test and eliminate the need for reference materials. As part of the overall test battery, Part III is challenging, but the choice of topics and their broad relationship to language teaching should create a fair test. Topics more closely related to the field might well create interference in the scoring. If the objective is to assign a writing proficiency level, the examinee's subject-matter expertise should not be evaluated. Applicants have other opportunities to present their professional credentials. On the other hand, examinees can define the topic as they choose. Their professional experience may well influence the organization and development of their ideas into a style appropriate to the audience.

These three diverse assignments should allow higher-level examinees to demonstrate their full proficiency. Both formal and informal styles are required. Topics are social and professional, affording three opportunities to develop and organize ideas and to demonstrate control of grammar and strength of vocabulary. A certain practicality is inherent in the tasks. Finally, the writer must tailor his or her presentation to three quite different audiences. Even the Level 5, who may not consider these topics especially stimulating intellectually, should find that he or she has been tested against the level descriptions within realistic time constraints.

These constraints will have to be determined through field trial. At this point, it appears that 2.5 hours should be allowed for completion of Parts I and II, followed by a break. Examinees who want to qualify only for the minimum score could leave at the break. This procedure would provide a thorough test of Level 2 without submitting lower-level writers to a frustrating experience. Those who return to complete Part III should be given no more than 2.5 hours.

Naturally, rigorous validation would be needed to ensure that these or other topics are appropriate and sufficient to discriminate at the required levels. In addition, multiple test forms should be prepared.

Following validation and trial implementation of the English proficiency test, similar but not identical target-language tests should be developed. Such a design could include a topic similar to Part II of the English test and designate a specific target-culture reader. Two separate assignments similar to those in Part III would be useful. These should address distinct audiences, with contrasting requirements for topic, tasks, and style; however, both should be appropriate for the target culture.

The instructions and the scoring criteria should be equally sensitive to rhetorical patterns in the language tested. In this respect, testing specialists have a great many questions to ask the native speakers/writers who are certified to test their language. While little research has been done on contrastive rhetoric, Kaplan (1966, 1972) certainly has raised important issues. The ILR descriptions and the ACTFL guidelines for writing must be considered tentative until his assumptions have been tested (see Gregg, 1986; Mohan, 1986a, 1986b; Mohan & Lo, 1985; Ricento, 1986). The DLI experience seems to support Kaplan's contention that English rhetorical conventions do not apply universally (see also Carrell, 1982; Carrell & Eisterhold, 1983). Until evidence appears to the contrary, an English model should not be imposed on the writing of other languages. Not only should topics be culturally oriented, but instructions to both examinee and raters should refer to the rhetorical features characteristic of good prose in the target language.

Scoring of the English prototype, at least initially, should follow the holistic method now used at DLI. Two certified testers rate the essay independently, consulting only the level descriptions. Discrepancies that cross a major scale border (e.g., 2^+ and 3) require involving a third rater to determine final scoring (see Lowe, 1978, 1982, for comparable discussions of the third rater's role in speaking tests).

Analysis of the Viability of the Scale

Working on an English prototype, trying to apply the design to a variety of languages, and trying to anticipate scoring difficulties led to doubts concerning the viability of the scale. The improvements begun by Lowe and Palmer are clear to those familiar with the earlier version.[4] The current writing scale is generally commensurate with the speaking scale in terms of language production, and with the reading scale in terms of text. However, it now appears that much has been omitted.

The writing descriptions stress sentence-level skills. Perhaps as an inheritance from the previous iteration of government descriptions, emphasis is placed on spelling, punctuation, and control of grammatical structure; lexical accuracy is considered more fully than style. Many of the factors traditionally considered when evaluating essays are minimized or eliminated. Examples include organization, methods of development, diction and tone, creativity (or the classical "invention"), and attention to the audience. (Weaver, 1967, and Brooks & Warren, 1972, present the fundamentals of rhetoric for the native writer.)

Although it was not their purpose, companion essays by Dvorak (1986) and Osterholm (1986) reinforce the impression that the writing descriptions have crucial pieces missing. Osterholm points out that focusing on lower-level goals effectively blocks beginning writers from achieving mid-level goals such as developing a paragraph or using the paragraph to support the central argument. Dvorak refers to this concern for the conventions of language form as "transcription" and uses the term "composition" for "the skills involved in effectively developing and communicating an idea or making a point" (p. 145). She adds that for the last 20 years, most articles on foreign language writing have concentrated on the lower-level skills, discussing ways to "reduce and repair error damage" (p. 148). Usually, foreign language teachers concentrate on transcription

and regard composition, as Dvorak defines it, as outside their scope of interest.

It was within this general atmosphere that successive government committees narrowed the scope of the writing proficiency descriptions. However, by following the comparatively recent trend of foreign language classroom evaluation and ignoring several centuries of rhetorical analysis, DLI has not only eliminated valuable tools for evaluating prose but, theoretically, may have contributed to the tendency of examinees to become blocked by sentence-level problems. The damage is no doubt indeed only "theoretical," thanks to the extremely limited use of the descriptions to date.

The ILR Testing Committee inadvertently allowed this to happen both because of the infrequent use of the writing scale and because of an assumption that the two productive skills were highly parallel. Of course, parallels do exist. However, the oral interview does not examine polished, practiced speech; in fact, considerable effort is made to prevent the examinee from using prepared material. While the examiners' activities are highly structured, the interview places the examinee in a series of situations that closely approximate spontaneous speech acts of daily life. Writing cannot be tested in this way. Little writing beyond phone messages and notes is spontaneous. Most writing is expected to be polished and organized for the reader. Academic assignments—tests, essays, term papers—and professional reports and papers all require planning, organization, and revision. Except for the necessary time limitations, a test of writing proficiency makes almost the same demands on the writer as the writing done for academic or professional purposes. A writing test is a performance test. Therefore, the scoring system should include the factors used in evaluating the external criterion—the successful essay, report, or paper.

A traditional rhetorical analysis of an English essay would place considerable emphasis on organization. Attention would be given to the writer's skill in ordering material in a way that fits the subject, limiting the

scope, providing the intended emphasis, creating effective transitions, and achieving overall unity and coherence within a framework that develops ideas clearly for the intended reader. These and other aspects of organization are so crucial to the evaluation of prose that it may seem that the writing proficiency descriptions refer to something else. In a sense, of course, they do; the level descriptions focus on a writer's performance across a variety of tasks rather than on the characteristics of a single essay. However, rating performance does not alter the fact that a proficiency level can be assigned only to a writer after evaluating examples of his or her prose. Whether traditional scoring is global or factored, organization must be seriously examined. An effective Level 3 description should define a type and degree of organization acceptable for communication and clarity in general professional writing. The higher-level descriptions should indicate progressive refinements in skill and greater sophistication in types of prose organization. Conversely, lower-level descriptions should suggest the limitations of the less proficient writer in ordering material and making transitions.

Unfortunately, when revising the writing descriptions, the ILR committee did not focus on this part of the task. Instead of defining Level 3 as the threshold of basic organizational competence, the ILR description states, "Relationship of ideas is consistently clear." This is accurate but not adequate. Level 3+ is defined negatively, omitting any reference to improvements over Level 3 proficiency: "Organization may suffer due to lack of variety in organizational patterns or in variety of cohesive devices." Level 4 is described with more precision, although, like the 3+ description, it may overemphasize variety: "Expository prose is clearly, consistently and explicitly organized. The writer employs a variety of organizational patterns, uses a wide variety of cohesive devices such as ellipsis and parallelisms, and subordination in a variety of ways." Organization is not mentioned at levels higher than 4. The lower-level descriptions provide a fairly good idea of organizational

limitations. The 0+ description makes it clear that nothing beyond lists and filling out forms can be expected. Level 1 states adequately, "Writing tends to be a loose collection of sentences (or fragments) on a given topic and provides little evidence of conscious organization." The description says the Level 1+ writer "can create sentences and short paragraphs" but "generally cannot use basic cohesive elements of discourse to advantage." The Level 2 description is cryptic—"Uses a limited number of cohesive devices"—and the Level 2+ description does not mention organization.

Methods of development at the essay and paragraph level are also intrinsic to traditional rhetoric. This factor is so fully covered in the descriptions of reading levels (with descriptive and narrative texts found at Level 2 and argumentation at Level 3) that one would expect to find parallel references in the writing descriptions. This is not the case.

On the other hand, the ACTFL guidelines do address discourse functions. "Narratives and descriptions of a factual nature" are mentioned in the Advanced-level guidelines; Advanced-Plus states, "can describe and narrate personal experiences fully but has difficulty supporting points of view in written discourse"; and the Superior-level writer is said to be able to "hypothesize and present arguments or points of view accurately and effectively." The government descriptions would benefit from incorporating this approach, although research may be needed to validate it.

The place of style in evaluating prose has not been completely neglected in the descriptions. There are references to diction in terms of control or precision of vocabulary, which may be the only point that low-proficiency writers can be expected to consider. The Level 2+ description touches on aspects of style, both positive and negative, that may emerge: "Often shows surprising fluency and ease of expression. . . . Normally controls general vocabulary with some misuse of everyday vocabulary evident. Shows a limited ability to use circumlocution. . . . Style is still obviously foreign."

The Level 3 description covers the topic in relatively abstract terms: "Able to use the language effectively in most formal and informal written exchanges on practical, social, and professional topics." Adequacy of vocabulary and the foreign character of style are also mentioned. The Level 3+ description refers to use of "a few prose styles" and notes possible weakness in expressing "subtleties and nuances." The higher-level descriptions continue at the same level of abstraction to suggest graduated control. For example, the Level 4 description states: "Able to write the language precisely and accurately in a variety of prose styles," and the Level 4+ description adds a reference to "use of stylistic devices."

Without doubt, the entire writing scale would be improved by analyses of the contribution of diction and tone to style, the role of figurative language, the expressive effect of sentences with various rhetorical patterns, and the writer's awareness of connotations of words. These facets of style distinguish the prose of higher-level writers, although research is clearly needed to determine the level at which sensitivity to connotation emerges or when figurative language can be integrated into a text. While stylistic concerns should not dominate the level descriptions, a more detailed discussion of style would increase their usefulness.

The aspect of traditional rhetorical analysis that is most conspicuously absent from the descriptions is invention or creativity. Nothing is said at the lower levels. The Level 4 description states that writing is "adequate to express all his/her experiences," and Level 5 says, "In addition to being clear, explicit and informative, the writing and ideas are also imaginative." While these brief statements imply a great deal, they do not constitute an adequate analysis of the topic, particularly given that educated adults with writing proficiency below Level 4 have such limited control of rhetorical elements in the second language that they cannot fully express their thoughts. The gradually emerging ability to break the barriers to invention should be traceable

with a certain amount of precision in the Levels 2 through 5.

Overall, the current level descriptions say too little about organization, methods of development, and invention. Style and a factor that may be called "attention to the audience" or "reader-based writing" are included, but not in useful proportions. Sentence-level matters—grammar, punctuation, spelling, and mechanics—are disproportionately emphasized.

These deficiencies do not make the current scale unusable. The employees whose writing samples are rated against this scale may be required to write correspondence in English for circulation within DLI as well as target-language material for beginning students. Skilled evaluators can probably rate their potential ability to perform these tasks and could do a better job with a more thorough test. Because the current descriptions discuss the quality of correspondence, reports, and job-related writing, they are certainly adequate for DLI's purpose. However, it would be worthwhile to bring them closer to the descriptions of the other skill modalities with a view toward broader application.

Value of an Improved Scale

An improved writing scale would be as important as the improved writing proficiency test discussed earlier. Descriptions more aligned with traditional rhetoric could be used to evaluate candidates for educational programs. Government employees would find it beneficial to cite their English writing proficiency score when applying for graduate school. Universities would probably find a Level 3 or 4 more meaningful if the description provided reference points for organization and methods of development rather than merely for sentence-level skills.

Another potential use of an improved writing scale

would be for accrediting translators. While prospective translators should be evaluated in several skills, in their case writing is certainly one of the most important. To be effective, the translator must be able to write competently at the level of the original text. A higher-level reader with appropriate subject matter and cultural knowledge, but without appropriate writing skills, may be able to perform several functions successfully. He or she may be able to prepare reasonably accurate summaries of a text. He or she may be able to prepare answers to specific questions that interest readers. However, unless he or she can replicate the author's presentation, convey subtleties, and address the audience with the tone of the originator, the translator will lose the qualities that make the text unique and thus worth reading. Those selecting translators for particular projects might well appreciate a writing proficiency score related to the aspects of texts that merit their translation.

A Method of Improving the Scale

The level descriptions are not to be rewritten here. The ILR experience has shown the value of developing the descriptions in a committee of interested people who have worked with the concepts extensively and gained insights that discussion will bring out. Judgments should not be hasty, for a great deal of research is needed.

Therefore, the following procedure is recommended. For several years, the CIA Language School and DLI have used a Speaking Performance Profile as part of the training for the oral interview; the Government Language School has integrated this aid into the scoring of interviews (see Lowe, 1982). The profile isolates six factors, all related to the speaking descriptions—pronunciation, fluency, sociolinguistic/cultural

information, grammar, vocabulary, and tasks—and charts each on a 0-to-5 scale with a brief shorthand description. (A copy is reproduced in the appendix.) The profile complements but neither duplicates nor substitutes for the full descriptions.[5]

An analogous profile should be constructed for writing, stating the performance characteristic of each level for six factors: organization, methods of development, style, invention, reader focus, and sentence-level skills. The initial statements would be tentative, based on DLI's empirical experience and collective judgment. DLI, and perhaps other government agencies, could use the profile together with the new writing proficiency test until enough samples and proficiency data are accumulated to test the hypothesis of the profile. Trial, observation, and concurrent research could then provide a basis for refining the profile and revising the level descriptions.

If a profile based on traditional rhetorical analysis seems a step backward to those interested in writing as process (see Osterholm, 1986), it can be argued that the tradition is useful and is more compatible with proficiency than the strictly sentence-level concerns that pervade the current descriptions. Nevertheless, the current proposal appears to emphasize writing-as-product more strongly than ever in the attention given to completed texts. Research into both areas is needed; and the results should be mutually beneficial. It is particularly important to determine the effect on a writer's proficiency when he or she encounters varying topics, under varying conditions, and whether tests with optional equivalent topics provide better opportunities to demonstrate writing skills. (See Coffman, 1971.) At some point, DLI evaluation must focus on the product created by the examinee. However, research into process should help in developing the best test to elicit those products. Traditional rhetoric should provide a better model for evaluation by putting all the factors that go into writing proficiency into proper proportion.

Nature of an Improved Scale

To show how the profile might look, Figures 5.2 and 5.3 show preliminary statements for two factors—organization and methods of development. These first drafts, based entirely on the author's experience, are intended to begin discussion. Colleagues in the government, and those in the academic community who have applied the ACTFL writing guidelines, may bring forward quite different observations to help clarify the content and condense the statements about these 11-point analyses.

0 No functional ability.

0+ Can produce lists and fill in blanks appropriately.

1 Writing tends to be a loose collection of sentences (or fragments) on a given topic and provides little evidence of conscious organization.

1+ Can produce short paragraphs on limited topics (e.g., survival and social needs). Generally cannot use basic cohesive elements of discourse to advantage.

2 Can produce a series of paragraphs on limited topics. Uses a limited number of cohesive devices and a kind of ordering of major points. However, lack of unity and proportion may confuse reader. Relationship of ideas may not always be clear. There are few meaningful transitions.

2+ Can produce a series of paragraphs that form a text; narratives and descriptions are usually organized adequately. Relationship of ideas is generally clear; there is evidence of effort to keep parts in proportion. Uses cohesive devices with some success. However, clarity, unity, and coherence may be flawed in any text.

3 Narratives and descriptions are well organized; expository prose is usually organized adequately. Ordering of major points fits purpose of the text. Material is presented coherently. Relationship of ideas is clear. Transitions are usually successful.

Figure 5.2. Profile of Writing Organization

3+ Usually able to fit the type of organization to subject matter and purpose. Narration, description, and exposition are always well organized. Can usually present arguments clearly and coherently. Texts are normally unified. Transitions are nearly always successful.

4 Can regularly organize various types of narrative, description, exposition, and argumentation appropriately. Even complex ideas and unique points of view can be ordered clearly. Transitions effectively aid understanding.

4+ Good ability to organize all types of writing according to needs of subject matter and purpose. Clarity is created through unity and coherence. Parts are always in proportion to whole. Transitions are skillful.

5 Strong ability to organize all types of writing according to needs of subject matter and purpose. Clarity is created through unity and coherence even for complex and unusual subjects. There is a strong effect of correct proportion. Transitions are skillful and well integrated into text. A degree of originality is often present in organizational techniques.

Figure 5.2. (cont.)

Despite attempts to keep the criteria broad, this preliminary version is biased toward English prose organization. A similar analysis of methods of development is even more obviously restricted to the principles that apply to English paragraphs and essays (see Carrell & Eisterhold, 1983). Considerable discussion and research will be needed to determine whether these conclusions have any application to other languages or whether more language-general statements can be framed.

The statements on methods of development are also briefer and more tentative than those on organization. A proficiency level has been assigned for the lowest point at which some practicable use may be made of a method; naturally, any method can be used with greater effectiveness at a higher level. Again, it should be clear that research is needed to determine the order of difficulty, to set cut-off levels or second-language writers' minimal control of a method, and to learn more about the language-specific issues. Because the subject of style is uncharted territory and far more complex than organization and methods of development, it

0	No functional ability.
0+	Can produce lists (on familiar topics).
1	Can produce statements (on familiar topics).
1+	Can produce simple time-ordered narration of anecdotes or incidents; simple descriptions of common objects, rooms, places.
2	Can use elementary classification (e.g., subjects in school, shops in mall); logical definitions.
2+	Can use exposition of process (e.g., how to change a tire or bake bread); more complex narration, including attention to narrator's perspective and conflict; example and illustration; comparison and contrast; cause and effect; narration that includes a sketch or profile; description that includes a variety of sensory impressions.
3	Can produce analysis that goes beyond classification, process, or comparison to present and support an opinion; arguments to support an opinion; elaborated descriptions (for example, showing group interaction).
3+	Can produce an extended definition (for purposes of persuasion); argumentation that is logically and rhetorically elaborated, combining many of the previously mentioned methods of development; persuasion involving ethical or philosophical arguments.
4	Can write professional or educational essays that successfully combine factual reporting or documentation with supported opinion, analysis, and other techniques. (Success implies ensuring that the reader can readily discern the difference between fact and opinion. Discourse development will be highly integrated with style and organization.)
4+/5	Controls full range of methods of development available to the educated native writer, including conventions applied in a specific field (e.g., law, medicine, science, military, literature, etc.).

Figure 5.3. Profile of Methods of Writing Development

---

would be premature even to begin a Profile until rated writing samples are gathered for research. Such samples should be analyzed from at least four perspectives.

Diction. The choice of both concrete and abstract words should be analyzed to determine the writer's understanding of their denotative and connotative value. Starting at about Level 2^+, the writer should show awareness of connotations. The use of technical vocabulary at Level 3 and above should also be examined.

Figurative Language. Considerable work needs to be done to learn how much use and control can be expected at Level 3 and above, assuming that little figurative language appears below that level. As a preliminary hypothesis, the order of usage may be as follows: imagery, simile, metaphor, allusion, symbol, and trope.

Tone. Attitudes toward the subject and toward the reader should begin to be controlled at Level 3 and above. Irony, understatement, overstatement, hyperbole, and humor should be available in the repertoire of writers at Level 3^+ and above. Since maladroit writing must also be rated, some attention should be given to the place of circumlocution, euphemism, and cliché in setting the tone.

Rhetorical Patterns of Sentence. The rhythm of prose for both formal and informal style should be considered at Level 3 and above.

In the attempts to add invention to the profile and, hence, to the level descriptions, analysis of writing samples will be needed. Perhaps, also, current studies of writing-as-process will provide useful information on the acquisition of this factor. As noted earlier, the current level descriptions include specific statements about reader focus and sentence-level skills; it should not be difficult to construct preliminary scales for these factors. There is no question that profiles would have to be tried and revised as lessons are learned from experience and research.

The Need for Research

This chapter concludes with a list of research topics that, ideally, would accompany the construction and trial of writing proficiency tests and collection of rated samples. Development of a Writing Performance Profile should occur concurrently, and eventually the completed profile, the test results, and research conclusions could all be brought together in a project to rewrite the level descriptions. Research needed to support this effort includes the following:

1. Determination of the validity and reliability of any essay examination as a performance test of writing proficiency—given time restrictions, the limited opportunity to revise and polish, and the relatively small number of topics offered to sample the skill modality. Determination must also be made of the external criterion for such a performance test and how to use it for validation (see Coffman, 1971).
2. Determination of the validity and reliability of the specific type of sampling proposed in this chapter and by Magnan (1985).
3. Determination of inter- and intrarater reliability of scoring such tests according to the ILR level descriptions and the ACTFL guidelines. The fundamental problems have been covered by Coffman (1971). Because of the inadequacies of the current DLI topic for evaluating the full 0-5 range, reliability statistics have not been collected for the scoring of writing. Informal observation suggests no greater need to use third raters for scoring essays than for speaking samples; interrater reliability for speaking tests has traditionally exceeded .80 at all government agencies.
4. Full-scale examination of contrastive rhetoric, as suggested by Kaplan (1972), to learn both about possible differences across languages and about the implications such differences would have for essay test design, topics, scoring criteria, and instructions to examinees.

If results suggest extensive differences, study and discussion would necessarily follow to assure that new level descriptions avoid bias toward more commonly taught languages.

5. Comparison of the proposed Writing Performance Profile factors to global ratings when evaluating samples.

6. Although the inherent philosophy of the level descriptions indicates a global score must be assigned, some consideration should be given to the inclusion of part scores for the proposed tests.

7. Examination of the role of writing proficiency as part of a translator's skill.

8. Research examining the process of writing by both natives and nonnatives in a variety of professional disciplines should be monitored for relevance to revised level descriptions.

9. Examination of native-writer norms. Two types of good writing can be postulated. One type is produced by artists or essayists whose imagination and originality distinguish them as more creative than their peers. The other type is produced by the educated writer who has something to say, organizes and states it well, but displays no real creative genius. Should only the former be eligible for a 4^+ or 5 rating? The results of all research on writing must be used to determine what constitutes successful writing by natives so that tests do not demand more from nonnatives.

NOTES

1. The Level 2 descriptions included the statement, "Material normally requires editing by a more proficient writer"; Level 3, "All formal writing needs to be edited by an educated native"; Level 4, "Errors are rare and do not interfere with understanding. Nevertheless, drafts of official correspondence and documents need to be edited by an educated native."

2. Without associating them with any shortcomings found in the prototype test, the author wishes to acknowledge the assistance of Anne Wright, Robert Kluender, and Carl Erickson. All were on the DLI staff at that time.

3. Magnan's test made it clear that evaluative information was an important addition to the instructions. The higher-level test was amended accordingly; all wording is derived from Magnan's test.

4. The addition of plus-level descriptions permitted more attention to the gradation of skills. For example, the current 0^+ description appropriately contains the statement, formerly found at Level 1: "Can write numbers and dates, own name, nationality, address, etc." References to the need for an editor's assistance at Level 4 and below have been removed, although the ability to edit texts is noted at Levels 4^+ and 5.

5. The rating is global; the profile factors are not totaled to yield a score.

References

Adams, M.L., & Frith, J. (Eds.) (1979). *Testing kit.* Arlington, VA: Foreign Service Institute.

Brooks, C., & Warren, R.P. (1972). *Modern rhetoric.* New York: Harcourt Brace Jovanovich.

Carrell, P.L. (1982). Cohesion is not coherence. *TESOL Quarterly, 16,* 479-488.

Carrell, P.L., & Eisterhold, J.C. (1983). Schema theory and ESL reading pedagogy. *TESOL Quarterly, 17,* 553-573.

Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.), *Educational measurement.* Washington, DC: American Council on Education.

Dvorak, T. (1986). Writing in the foreign language. In B.H. Wing (Ed.), *Listening, reading, writing:*

Analysis and application (Reports of the Northeast Conference on the Teaching of Foreign Languages, pp. 145-167). Middlebury, VT: Northeast Conference.

Gregg, J. (1986). Comments on Bernard A. Mohan and Winnie Au-Yeung Lo's "Academic writing and Chinese students: Transfer and developmental factors." A reader reacts. *TESOL Quarterly, 20,* 354-358.

Kaplan, R.B. (1966). Cultural thought patterns in intercultural education. *Language Learning, 16,* 1-20.

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

Lowe, P., Jr. (1978). Third rating of FSI interviews. In J.L.D. Clark (Ed.), *Direct testing of speaking proficiency: Theory and application* (pp. 159-163). Princeton, NJ: Educational Testing Service.

Lowe, P., Jr. (1982). *ILR handbook on oral interview testing* (DLI/LS Joint Oral Interview Project). Arlington, VA: Foreign Service Institute Language School.

Magnan, S.S. (1985). Teaching and testing proficiency in writing: Skills to transcend the second-language classroom. In A.C. Omaggio (Ed.), *Proficiency, curriculum, articulation: The ties that bind* (Reports of the Northeast Conference on the Teaching of Foreign Languages, pp. 109-136). Middlebury, VT: Northeast Conference.

Mohan, B.A., & Lo, W.A. (1985). Academic writing and Chinese students: Transfer and developmental factors. *TESOL Quarterly, 19,* 515-534.

Mohan, B.A. (1986a). On evidence for cross-cultural rhetoric. *TESOL Quarterly, 20,* 358-361.

Mohan, B.A. (1986b). On hypotheses in cross-cultural rhetoric research. *TESOL Quarterly, 20,* 569-573.

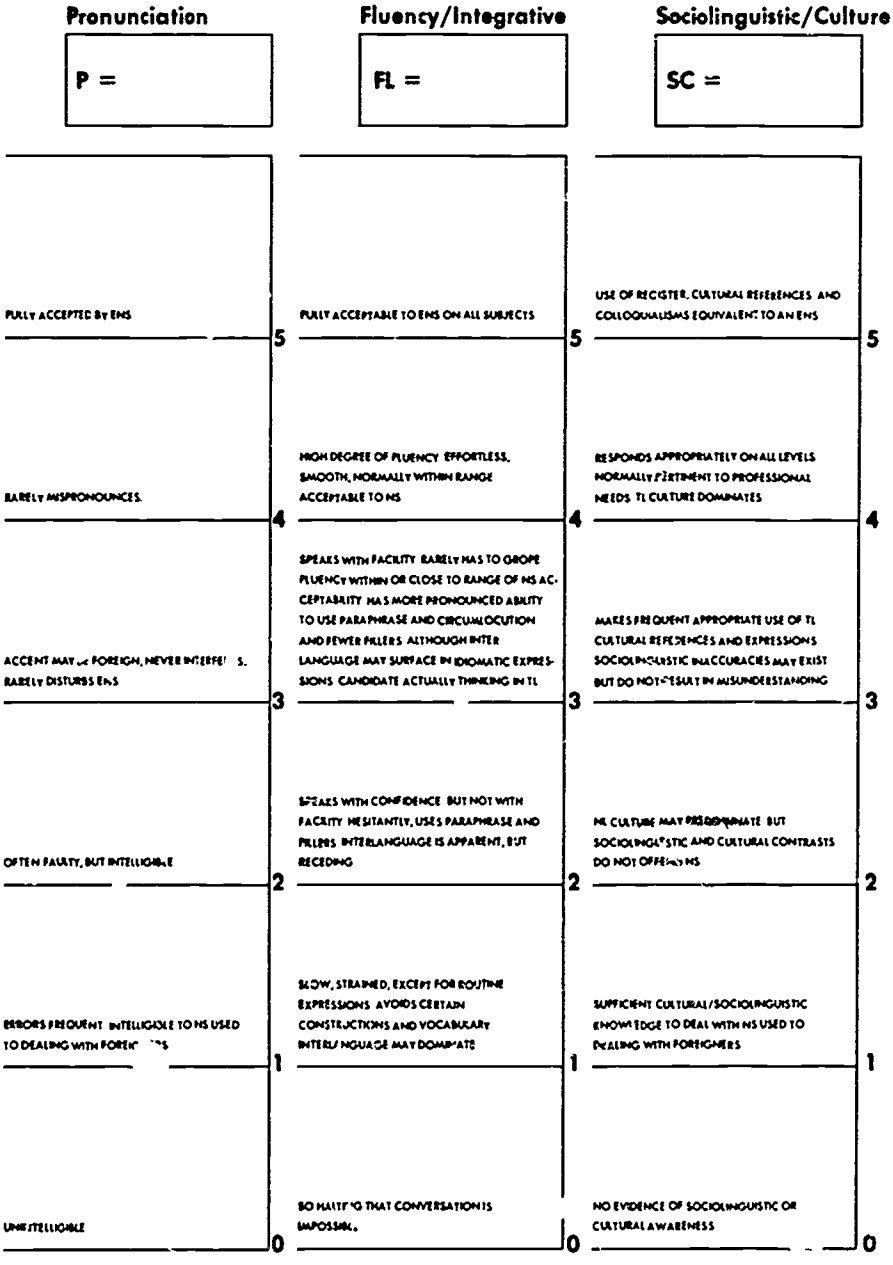
Osterholm, K. (1986). Writing in the native language. In B.H. Wing (Ed.), *Listening, reading, writing: Analysis and application* (Reports of the Northeast Conference on the Teaching of Foreign Languages, pp. 117-143). Middlebury, VT: Northeast Conference.

Ricento, T. (1986). Comments on Bernard A. Mohan and Winnie Au-Yeung Lo's "Academic writing and Chinese students: Transfer and developmental factors." *TESOL Quarterly, 20,* 565-568.

Sollenberger, H.E. (1978). Development and current use of the FSI oral interview test. In J.L.D. Clark (Ed.), *Direct testing of speaking proficiency: Theory and application* (pp. 1-12). Princeton, NJ: Educational Testing Service.

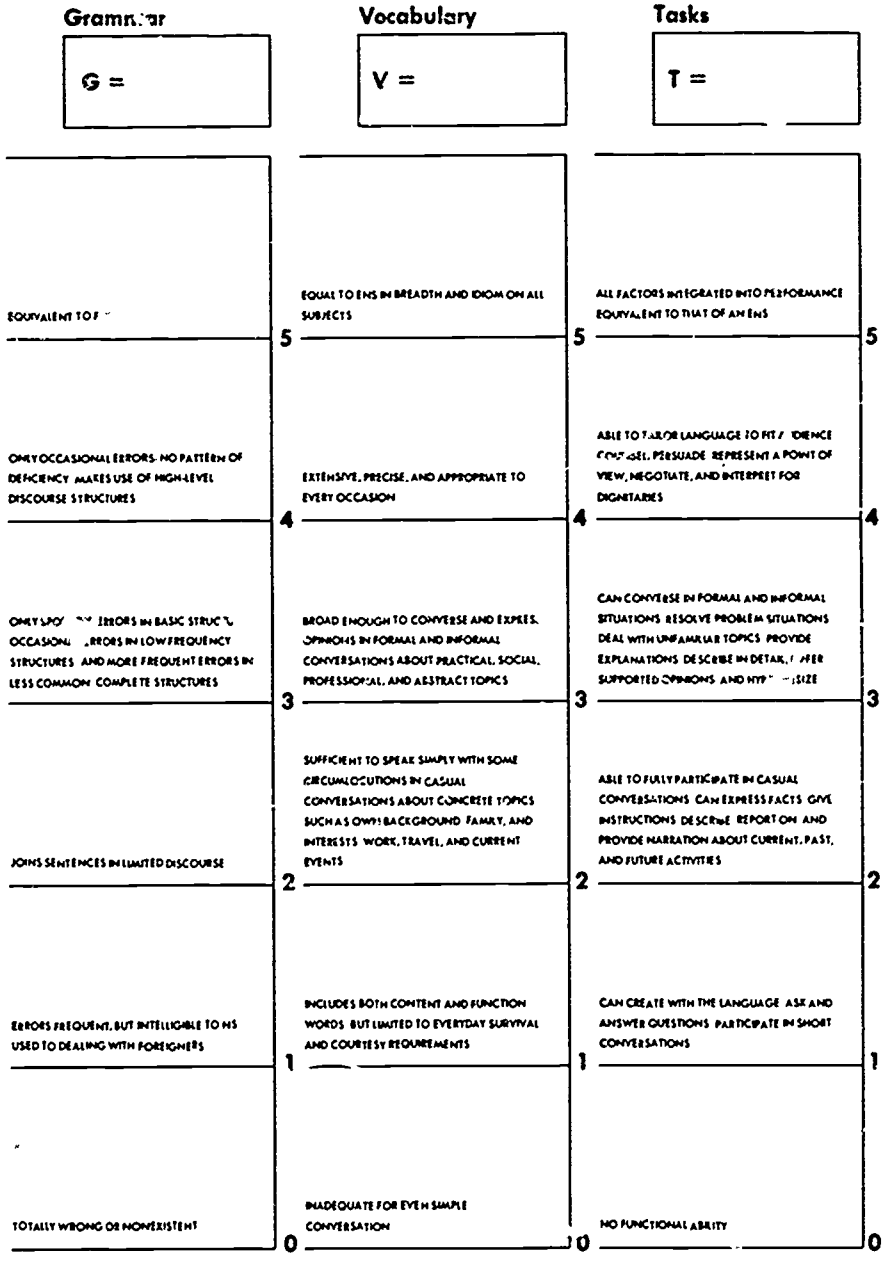
Weaver, R.M. (1967). *A rhetoric and handbook.* New York: Holt, Rinehart and Winston.

Appendix: Speaking Performance Profile



ENS = educated native speaker

NS = native speaker



TL = target language NL = native language

Section 2: The Academic Context

by Anne Katz,
University of California at Berkeley

Until recently, writing has been the neglected sister of the four language skill areas customarily taught in second and foreign language classes—listening, speaking, reading, and writing. During the long period when a traditional grammar approach dominated language teaching, writing was viewed as a technique to practice, and thus learn, grammar and the rules of correct usage. Tasks included copying words and sentences, dictation, and translating. The aim of this approach to writing was correctness of linguistic form, usually at the sentence level.

The postwar advent of the audio-lingual approach brought a major change in focus for many language programs, from a formal undertaking to a practical means of teaching oral communication. This new approach emphasized oral skills, with students practicing and learning the patterns of spoken language. Techniques for teaching writing remained essentially the same, however. Tasks still included copying, writing from dictation, and practicing word and sentence patterns. The aim of these tasks was to reinforce language patterns introduced in oral practice.

The advent of the proficiency movement, along with a resurgence of interest in developing students' writing skills, has opened up the possibility of viewing writing in the foreign language classroom in a new light (Magnan, 1985). With a focus on communicative approaches to teaching language, teachers have become more aware than ever of the importance of gauging students' needs in designing appropriate language teaching

curricula (Savignon, 1983). Writing research has pointed out the cognitive benefits accruing from using writing to discover and explore meaning in both language classes and content classes (Applebee, 1984).

Taking heed of this new perspective, some second and foreign language specialists interested in teaching writing have leaned heavily on the first language (L1) writing literature to provide both theory and techniques for helping student writers (for a sampling from both first and second language writing specialists, see Dvorak, 1986; Hughey, Wormuth, Hartfiel, Jacobs, 1983; Magnan, 1985; Osterholm, 1986; and Zamel, 1976, 1982, 1985). Drawing from this literature, second and foreign language writing specialists have redefined the province of language teachers, encouraging them to approach the teaching of writing as *process* rather than as *product*, from the *composition* side rather than from the *grammar* side. Thus, writing tasks are seen as ways to use language to communicate meaning rather than as ways to practice or test writing skills.

Uses of Writing in the Foreign Language Classroom

While descriptions of this new approach to teaching writing are surfacing in the pedagogical literature, a few studies of the use of proficiency procedures in the academic classroom are also emerging. The news, for those interested in writing, offers mixed results.

In a preliminary analysis of the effects of the proficiency-based foreign language requirement instituted at the University of Pennsylvania, Freed (1987) points out that the program is designed to help students "acquire functionally useful communicative abilities as well as structural accuracy in the basic four skills" (p. 140). To satisfy the requisite foreign language requirement, students must pass a proficiency test that includes exams

of functional skills in oral interaction, listening comprehension, and reading comprehension, as well as in writing.

This focus on multiskill proficiency has led to change in the curriculum, to what Freed describes as "more creative teaching and more emphasis on extended use of the language in all four skills" (p. 145). For teaching writing, this has meant "more varied and practical types of writing assignments, and a decreased emphasis on formal grammatical manipulation" (p. 142).

On the other hand, the proficiency movement's effect on other classrooms has produced somewhat different results. In describing an alternative first-year sequence in French, German, and Spanish, Clausen (1986) outlines a program that focuses on listening comprehension, culture, and speaking in everyday situations.

This innovative program is intended as an alternative to standard, grammar-oriented first-year classes. Clausen explains that her department is still working out how to implement a proficiency-oriented curriculum. While oral testing has been introduced, changes involving the other skill areas are "more sporadic and vary greatly from one instructor to another" (p. 35).

Although more uniform enthusiasm for including writing in the curriculum might be expected or desired, it is understandable that writing continues to play a secondary role. For while the proficiency orientation has provided new ways of thinking about language and has developed guidelines for all four skill areas, the emphasis in the classroom continues to be on oral fluency. As Dvorak (1986) notes, writing in the foreign language classroom is still widely regarded as "speech from a pencil" (p. 147).

One reason for the continued emphasis on oral fluency certainly is tied to the most recent history of language teaching methodology, as discussed earlier. The focus on proficiency and new ways of thinking about writing are relatively recent developments. Classroom

teachers have been trained according to principles embodied in previous approaches to language teaching. In reviewing modern methodologists of the late 1970s and early 1980s, Dvorak reports that their discussions of "writing" focus specifically on teaching the conventions of language form. For the most part, then, the development of "writing" skill is narrowly defined in terms of the development of language.

Dvorak also points out that even the textbooks students encounter in class continue to support this concept of writing. Written assignments are linked to conversation practice or advanced grammar lessons. In this way, writing can be seen essentially as transcription, as a means to extend the lesson or to vary the day's activities. "Advanced composition" is defined as either free composition, translation, or some combination of the two. Writing is evaluated according to the frequency and gravity of error in the student's product. From this perspective, the development of writing skill entails increased fluency and accuracy in the target language but says nothing about how well the learner has managed to deal with the content or with the notions of audience or purpose contained in the task.

While part of the difficulty in introducing new ways of thinking about writing into the foreign language classroom may stem from the traditional ways in which foreign language has been taught, it may also have something to do with the very nature of foreign language requirements. Few schools require extensive foreign language study. One year, at most two, will satisfy most university requirements. Freed (1987) notes, for example, that while the proficiency test is given to fourth-semester language students, students who wish to take it during the third semester may do so. Given the limited amount of time allotted for instruction, teachers and program coordinators generally give writing short shrift.

Writing as a means of conveying meaning is not normally a part of the curriculum except in upper-division or graduate-level courses in literature. Even

then, students are assumed to have mastered in other classes the necessary prerequisite skills of organizing and developing their ideas.

It is not hard to understand the rationale behind the role of writing in the foreign language classroom. One important factor concerns the generally perceived nature of language study. One learns a foreign language in order to speak it, not write it. Thus, when students enter the classroom, expecting to develop their communicative abilities, they are primarily expecting to learn to speak with other users of the foreign language. The strength of students' perceived needs is an important factor that is given a great deal of attention both in second language acquisition research (Ellis, 1985) and in syllabus design (Richards, 1984).

Writing is also perceived as the most difficult of the language skills, both by teachers and by students. To begin with, students often have difficulty learning how to transcribe units of meaning in the new language, and how to untangle a new system of sound-symbol correspondences. And while it may be difficult to detect whether inflectional endings are present in the flow of speech, there is no doubt as to their presence, or absence, when students begin writing a dictation or an essay. When writing is used in the classroom to extend the grammar lesson and vary the range of activities students engage in, and is evaluated in terms of accuracy and fluency instead of in terms of a process through which to capture meaning, these perceptions of difficulty are reinforced.

The Role of Proficiency Testing in the Foreign Language Classroom

It is not unknown in education for both teachers and students to work toward the test that will be used to evaluate the efforts of each (Purnell, 1982). Thus, a key

role that a test of writing proficiency may play in foreign language classrooms is that of spurring curricular innovation. It is crucial, then, to weigh the direction that a test of writing proficiency will take. Given the influence of the L1 writing literature on the emerging theory and practice of second and foreign language writing classrooms, it makes sense to tap into this rich vein for insights useful for shaping the direction needed in using the ACTFL guidelines for assessing writing proficiency.

Assessment of Writing Proficiency in English

Large-scale assessment of writing has emerged as a major trend in the field of writing today (Wolcott, 1987). While writing teachers and researchers explore and describe developing "processes" of students, it is students' "products" that continue to serve as indicators of their competence at critical points along the way. High schools, colleges, and universities require such information to make placement decisions and to determine more realistic levels of foreign language skills as students emerge from language programs to take their places in government, business, and industry (Omaggio, 1973). These data also serve to inform judgments about the worth of particular programs and to aid in making budget decisions.

Most of the literature on writing assessment revolves around assessing writing proficiency in English as a first language. Odell (1981) suggests the first step in assessing writing is to determine what is meant by competence in writing. He points out that such a notion must encompass a range of skills, including lexical, syntactic, and creative fluencies, discourse skills, and appreciational skills. His notion of competence also includes discovering what one wants to say in writing;

it requires the writer to come to some conclusion about the topic at hand. This aspect of competence involves the writer in making "appropriate" choices guided by an awareness of audience and purpose. This definition of competence makes it quite clear that different writing tasks require different writing skills, evoking a different balance of options to satisfy the demands of audience and purpose. Such a notion of writing competence entails a vision of writing as communication, not merely as a channel.

It also poses certain difficulties in devising appropriate tasks for assessing that competence. To measure competence, Odell underlines the importance of obtaining an adequate sample of students' writing and choosing an appropriate measure of writing competence.

According to Odell, obtaining an adequate sample entails devising assessment procedures that meet several criteria:

1. Students should be asked to write under conditions that resemble as closely as possible those under which "important" writing is done. To best represent their competence, students need time to "engage in the process of discovery."

2. Students should produce more than one type of writing. The rationale for this criterion is based on evidence that the "ability to do one kind of writing task may not imply equal ability with other kinds of tasks. As well, one kind of writing may not be equally important for all students in all schools in all communities." In addition to producing more than one type, students should be asked to write for more than one audience and for more than one purpose.

3. Information about the audience and the purpose of the piece of writing students produce should be included in the prompts for writing. Also, the prompts might indicate the form of response desired, be it a letter, a summary, an essay, or a journal entry.

4. The demands of the writing tasks assigned should be carefully assessed to determine whether different

topics require students to draw on different sources of development or to provide different modes of development.

5. Several pieces of writing (at least two for each kind of writing) should be collected to make a judgment of competency.

Once the samples of writing have been collected, it is necessary to choose an appropriate measure of writing competence. While objective, multiple-choice tests have long been used for determining competence in writing, one of the most common current means of assessing writing competence is holistic evaluation.

Holistic evaluation is a guided procedure for assessing a sample of writing. Cooper (1977) describes the procedure as follows:

> The rater takes a piece of writing and either (1) matches it with another piece in a graded series of pieces or (2) scores it for the prominence of certain features important to that kind of writing or (3) assigns it a letter or number grade. The placing, scoring, or grading occurs quickly, impressionistically, after the rater has practiced the procedure with other raters. (p. 3)

The strength of holistic evaluation lies in the apparent validity of its results. For, as Cooper continues:

> A piece of writing communicates a whole message with a particular tone to a known audience for some purpose: information, argument, amusement, ridicule, titillation. At present, holistic evaluation by a human respondent gets us closer to what is essential in such a communication than frequency counts do. (p. 3)

It is important, however, to remember that holistic evaluation is a tool, a procedure for assessing a piece of writing according to prescribed criteria. The strength of

the claim for validity rests in the set of features selected by the test designers. The features deemed essential by one group of readers may not necessarily be considered essential by another. In a critical overview of holistic scoring, Charney (1984) points out that deciding on the boundaries of categories for the purpose of evaluation or determining the salient features to be included in scoring guides are all "matters of opinion."

Charney's caveat has critical implications for applying the guidelines to create tests of writing, particularly across languages where different features may hold varying degrees of saliency. Native users from different rhetorical traditions may choose contrastive sets of features to characterize "good" writing.

A Specific Example of an L2 Rating Scale for Writing: The TOEFL Test of Written English

Before considering exactly how a model test of foreign language writing proficiency might look, it will be useful to examine an existing test of writing for nonnative users of the target language. One such major test is the new writing section of the Test of English as a Foreign Language (TOEFL), the Test of Written English (TWE).

The first issue facing the developers of a test of writing in English as a second language (ESL) was to redefine the notion of written competence. As Odell (1981) cautions, it is necessary to begin with an understanding of what it is being measured. In a discussion of the research framing the development of the TWE, Carlson and Bridgeman (1986) explain how they approached their initial research task from the perspective of "functional communicative competency." Contrasting their approach with the more traditional grammatically

based one, they defined a functionally based communicative approach as entailing "the ability to use language to communicate effectively within the specific context in which the communication takes place" (p. 129). The key assumption here is that performance on a task is conditioned by the specific dimensions of that task. To evaluate performance, then, those specific dimensions must be clarified.

Because the TOEFL is used as an admissions instrument for colleges and universities Carlson and Bridgeman designed a survey to assess the parameters of the tasks students face in an academic setting. Using the results of this survey, they devised an examination centered essentially on two very different types of writing: (a) a description and interpretation of a graph or chart, and (b) comparison and contrast, plus taking a position. Subjects wrote two essays of each type and were allotted 30 minutes to respond to each question.

These were the two comparison-and-contrast topics used in developing the test:

> Some people say that exploration of outer space has many advantages; other people feel that it is a waste of money and other resources. Write a brief essay in which you discuss each of these positions. Give one or two advantages and disadvantages of space exploration, and explain which position you support.
>
> Many people enjoy active physical recreation like sports and other forms of exercise. Other people prefer intellectual activities like reading or listening to music. In a brief essay, discuss one or two benefits of physical activities and of intellectual activities. Explain which kind of recreation you think is more valuable to someone your age.

For the first of two topics that required the description and interpretation of a chart or graph, students studied three graphs showing changes in farming in

the United States from 1940-1980 and followed these instructions:

> Suppose that you are writing a report in which you must interpret the three graphs shown above. Write the section of that report in which you discuss how the graphs are related to each other and explain the conclusions you have reached from the information in the graphs. Be sure the graphs support your conclusions.

The next graphic was a pair of charts showing the area and population of the continents, with these instructions:

> Suppose you are to write a report in which you interpret these charts. Discuss how the information in the Area chart is related to the information in the Population chart. Explain the conclusions you have reached from the information in the two charts. Be sure the charts support your conclusions.

These sample topics illustrate the kind of writing prompts TWE has continued to use since its first administration in July 1986. Overall, Carlson and Bridgeman found that with careful topic selection and adequate training of raters, they could reliably evaluate the writing performance of ESL students.

To score the essays produced in response to these prompts, Stansfield (1986) directed the development of a six-point scoring guide for use by raters trained in holistic evaluation of writing. The scoring guide is given in Figure 5.5.

While the defining criteria of each level cut clearly across the full range of writing performance, a range of ability remains within each of the six levels. Thus, for example, there are "high" Level 4s, papers that strain the upper boundary of minimal competence, as well as "low" Level 4s, papers that barely constitute the same designation. Various factors contribute to the range

6 Clearly demonstrates competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors. A paper in this category:

- is well organized and well developed
- effectively addresses the writing task
- uses appropriate details to support a thesis or illustrate ideas
- shows unity, coherence, and progression
- displays consistent facility in the use of language
- demonstrates syntactic variety and appropriate word choice

5 Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will have occasional errors. A paper in this category:

- is generally well organized and well developed, though it may have fewer details than does a 6 paper
- may address some parts of the task more effectively than others
- shows unity, coherence, and progression
- demonstrates some syntactic variety and range of vocabulary
- displays facility in language, though it may have more errors than does a 6 paper

4 Demonstrates minimal competence in writing on both the rhetorical and syntactic levels. A paper in this category:

- is adequately organized
- addresses the writing topic adequately but may slight parts of the task
- uses some details to support a thesis or illustrate ideas
- demonstrates adequate but undistinguished or inconsistent facility with syntax and usage
- may contain some serious errors that occasionally obscure meaning

3 Demonstrates some developing competence in writing, but it remains flawed on either the rhetorical or syntactic level, or both. A paper in this category may reveal one or more of the following weaknesses:

- inadequate organization or development
- failure to support or illustrate generalizations with appropriate or sufficient detail
- an accumulation of errors in sentence structure and/or usage
- a noticeably inappropriate choice of words or word forms

Figure 5.5. The TWE's 6-Point Guidelines

2 Suggests incompetence in writing. A paper in this category is seriously flawed by one or more of the following weaknesses:
- failure to organize or develop
- little or no detail, or irrelevant specifics
- serious and frequent errors in usage or sentence structure
- serious problems with focus

1 Demonstrates incompetence in writing. A paper in this category will contain serious and persistent writing errors, may be illogical or incoherent, or may reveal the writer's inability to comprehend the question. A paper that is severely underdeveloped also falls into this category.

Figure 5.5. (cont.)

of ability within levels, including the writer's skill in dealing with the topic or ability in selecting relevant detail to develop main points.

For ESL students, a major issue concerns the balance between rhetorical and syntactic skills. Since the scoring guidelines require testers to consider both rhetorical and syntactic skill when determining scores, essays exhibiting differing levels in each of these two areas could receive the same score because the *overall* effect of the essays could demonstrate similar levels of writing performance.

Carlson and Bridgeman point out that for native speakers, organization skills tend to parallel mechanical ones. With ESL students, however, greater disparity is to be expected. Carlson and Bridgeman suggest resolving this difficulty by having the readers reach a consensus about how to evaluate such essays. Clearly, the difficulties involved in dealing with a range of ability within levels are resolvable, as Carlson and Bridgeman reported consistently high reliability for the holistic scores (.80-.85, according to Spearman-Brown calculations of the reliability of a score based on the ratings).

The implementation of the TWE shows that it is possible to assess reliably the writing of students using their second or foreign language. It is now time to consider taking advantage of the insights garnered from the L1 writing assessment literature and from the implementation of the TWE to evaluate the proficiency

definitions and procedures for testing writing proficiency in a foreign language.

Assessing Foreign Language Proficiency

The first step in assessing writing, according to Odell (1981), is to determine what is meant by competence in writing. The proficiency guidelines determine how this competence will be defined. The four basic levels range from Novice, the ability to produce isolated words and phrases, to Superior, the ability to write formally and informally on practical, social, and professional topics. Within each level, the guidelines describe specific functions, content and degrees of accuracy appropriate to that level (see Magnan, 1985, for a fuller analysis of the levels organized according to these three areas).

One problem with the guidelines may be the high degree of specificity included for each level. For example, to receive a rating of Intermediate-Mid, a writer must be able to control the syntax of noncomplex sentences and basic inflectional morphology. The guidelines are presented as generic descriptions, yet second and foreign language specialists have collected little evidence to determine how well these descriptions work in assessing proficiency across the broad range of targeted languages (see Rosengrant, 1987, however, for a study of emerging grammatical and functional ability among Russian learners). Given Charney's caveat about the relative nature of categories used for evaluation, the descriptions should be considered tentative in nature. While specific criteria provide a clear focus for evaluators undertaking the task of assessing written proficiency, more cross-cultural research is needed to confirm the selection of features deemed characteristic of each level.

Another issue related to the guidelines' level of

specificity is the difficulty of tying specific sentence-level skills to rhetorical forms. It is necessary to consider how proficiency should be weighted in each of these areas. For, as Carlson and Bridgeman point out in their discussion of the TWE, greater disparity is expected in the growth of fluency in these skills among nonnative users of a language. The problem in assessment revolves around deriving a single score from unmatched skill levels.

Another and perhaps more disturbing problem with the guidelines concerns the linear model of development assumed across the levels of emerging proficiency. According to the guidelines, as students progress from Novice to Intermediate and beyond, they add new skills—students who acquire the ability to meet a number of practical writing needs (Intermediate-Mid level), for example, are still assumed to be able to supply information on simple forms and documents (Novice-High level). Within such a model, then, writers accrue skills, building on past successes with written language to forge texts containing newer, more complex linguistic structures and rhetorical functions.

Several difficulties emerge from this view of writing development. A linear model of development focuses attention on products—on the texts created by student writers—because under such a model, the evaluation of writing skill limits itself to examining the texts against specific criteria. Yet current writing research and instruction have reconceptualized what writing entails, enlarging the focus of inquiry to examine the cognitive processes writers engage in during the creation of written text. Researchers have suggested that an analysis of changes in written products may not reveal how (or perhaps even whether) writers have changed the way they go about producing text (Bereiter, 1980; Shuy, 1981). With this new focus on not only what student writers produce but also how they go about producing it, writing specialists are urging curricular and assessment innovations that take into account the process of writing. If writing proficiency is envisioned to include

these complex cognitive processes, then the question becomes how an assessment model that ostensibly focuses on products will take into consideration the underlying processes used to produce those texts.

Once the scope of the notion of writing is enlarged to include the variety of factors involved in the writing process—the demands of topic, audience, and purpose, for example—additional difficulties arise with a linear model of writing growth. In their study of indices of growth in writing, Freedman and Pringle (1980) examined the relationship between growth in rhetorical control and growth in the writer's capacity for abstraction to higher levels. They found that as students tackled more difficult topics or attempted more sophisticated lines of argument, their written texts did not appear as proficient as when they dealt with less challenging tasks. They argued that when students take on a more cognitively difficult task, they tax their rhetorical skills, with the result that their texts appear less skillful rhetorically and grammatically. "It is," they conclude, "quite simply more difficult to write when the task is more intellectually taxing" (p. 322).

Part of the difficulty in responding to more challenging tasks may have to do with the writer's level of knowledge about the topic. In discussing how writers interpreted writing tasks, Ruth and Murphy (1984) presented a complex model of how different background knowledge results in different interpretations of the task. For a striking example of the ambiguity residing in the interaction between a question and its receiver, they drew on the Interviewer's Manual from the Michigan Survey Research Center. When an interviewer asked, "Do you think the government should control profits or not?" the respondent answered, "Certainly not. Only Heaven should control prophets." Ruth and Murphy argued that such examples illustrate the illusion created by assuming that questioners and respondents, as well as test designers, takers, and raters, have a "common linguistic and social frame of reference." Given such potential complexity of interpretation, a

model of writing assessment, they suggest, should accommodate a range of possible responses.

While more discussion and research (for example, see Herzog's suggestions, this volume) will lead to fuller specifications of competence in the definitions to be used in evaluating the skills students develop, it is important now to consider the forms that evaluation can take in the classroom. As it was argued earlier, it is often the forms of assessment that drive, or at least reinforce, the foci of instruction in the classroom. AEI testing procedures must be devised that take into account the curricular innovations that foreign language writing specialists have suggested and that satisify the requirements of programmatic assessment.

The enormous range of the AEI guidelines precludes extensive discussion of all the various forms of assessment needed to determine foreign language writing proficiency. In the academic community, however, the skills attained in an intermediate-level course are often designated as the criterion level for satisfying foreign language requirements. Given the kinds of writing described for the Intermediate-High through Advanced Levels, the following discussion provides some suggestions for testing writing proficiency.

Holistic Evaluation

Explicit in the guidelines themselves is a holistic criterion. Lowe (1986) describes one of the guidelines' fixed characteristics as "expression of ability in a global rating" (p. 394). Thus, even though each level of the guidelines describes specific, required abilities, evaluation requires raters to take all of these "parts" and blend them into a "whole" score. This key characteristic fits in nicely with the tenets of holistic evaluation described earlier.

A characteristic of holistic evaluation that has been described as a limitation may turn out to be a strength

in the context of proficiency assessment. To achieve reliability, raters often need to interpret and negotiate the criteria used for scoring (a procedure described, for example, by Carlson & Bridgeman, 1986). Sample papers are used to flesh out the guidelines and provide a basis for such negotiation of meaning. While this discussion may be said to create an artificial aura of reliability (Charney, 1984), it also creates a community of readers with shared criteria and, thus, acknowledges the interaction that occurs between text and reader. For in evaluating a sample of text, raters draw not only on the surface features of the sample of writing, but also on their implicit notions about textual conventions. Discussions about samples of writing evince these implicit notions, allowing raters to discard or retain them as the training sessions reach a consensus.

Because the guidelines have sparked a variety of responses from foreign language practitioners who believe they do not address the full range of ability found in foreign language classrooms, the negotiation inherent in implementing the guidelines may provide a way for disparate viewpoints to use the scale. One example of a successful melding of two viewpoints through implementing the guidelines is reported by Hoffman and James (1986). They describe the ACTFL proficiency rating system as a "splendid tool" for integrating the foreign language and literature foci in their department. By designing a range of assignments in a literary context, they found that they could set intellectually rigorous literary tasks for all students, notwithstanding their varying foreign language proficiencies. Thus, they found that assessment was not limited to language or literary criteria alone, but combined both foci.

Given the leeway for negotiation, it would make sense, then, to devise assessment procedures using a holistic form of evaluation. Holistic evaluation, however, is a tool when used within an assessment plan. In the TWE, for example, holistic evaluation is used to assess one sample of writing from each candidate. This makes sense given that the TWE is designed as a test of

writing *performance* within predetermined types of writing. Generally, for reasons of financial economy, holistic evaluation in large assessment situations has been used to assess only one piece of writing per writer.

Yet limiting evaluation to one sample violates many of the tenets of "good" assessment discussed in the earlier review of L1 writing assessment. Lloyd-Jones (1982) goes so far as to state, "A writing sample is not real writing" (p. 3). Obtaining only one sample would also seem to violate the intent of the guidelines themselves. The guidelines are written as descriptions of *proficiency*. As such, they describe what the language learner is "able" to do at each level. Given what is known about variability in writing resulting from differences in type of writing, perceived audience, wording of topic, and time allowed for writing, it would be difficult to assess "ability" from one piece of writing. Thus assessment should not focus on a single product of specific conditions, but rather must include a variety of tasks, situations, types, and audiences (see Herzog, this volume). To assess learners' writing ability adequately requires a different approach to AEI assessment.

Portfolio Assessment

Portfolio assessment is a method by which students enrolled in a writing course produce a collection or "portfolio" of pieces of writing to be evaluated by a teacher other than the classroom teacher. The portfolios contain several types of writing and include both impromptu and revised papers (Elbow, 1986). Generally, the writing is evaluated holistically, although some pieces may be evaluated using other scoring methods.

The key advantage of this method is that it allows raters to evaluate a variety of types of writing. The portfolio may contain samples of different modes of dis-

course. Pieces of writing may be designed for different purposes and different audiences. This advantage ties in with Odell's argument that to evaluate students' ability, "we need to evaluate several different kinds of writing performance" (p. 115). Van Oorsouw (1986) points out that in addition to allowing a fairer assessment of students' ability, "it also offers the evaluators the opportunity to learn much more about their students' writing than other methods offer" (p. 20).

This method also allows students to develop a more realistic sense of who is assessing the writing and what the criteria are for those judgments. Hoetker (1982) suggests that while elaborate fictional topics may work well in teaching writing, they do not serve student writers as well in assessment situations. Rather, he suggests that a set of instructions describing the actual rhetorical context might provide students with useful information. Portfolio assessment provides such rhetorical context.

This method also draws the evaluators together, leading to discussion about criteria and what goes on in the classroom. As in the description of holistic scoring, portfolio assessment engenders a sense of community among the evaluators as they determine the standards to be implemented.

Finally, this method allows students to draw on the skills they learn in process-centered classrooms. Students are evaluated on pieces of writing that they can plan and revise, and that take time to produce. The method works in both directions, for the criteria used to evaluate the writing will be brought back to the classroom.

While the benefits of portfolio assessment are great, there are costs. Obviously, this method of evaluation involves considerable planning by both teachers and administrators. Because of increased costs in processing more papers for evaluation, the net bottom line will be greater. In addition, the method works best when it is part of an academic course. Large-scale assessment for placement purposes, for example, would be difficult to organize and carry out. Despite the difficulties, portfolio

assessment offers useful possibilities for evaluating AEI foreign language proficiency.

To determine future directions for both defining proficiency levels and testing them, further research is needed on how students develop proficiency in foreign languages. How closely are students' writing skills tied to reading skills? Do students progress in similar ways across different cultures? Do different languages and cultures place the same value on various types of discourse as English? The answers to these questions would provide needed direction for further changes in the proficiency guidelines and may prompt new thinking about writing and its processes.

References

Applebee, A.N. (1984). *Contexts for learning to write: Studies of secondary school instruction.* Norwood, NJ: Ablex.

Bereiter, C. (1980). Development in writing. In L.W. Gregg & E.R. Steinberg (Eds.), *Cognitive processes in writing.* Hillsdale, NJ: Erlbaum.

Carlson, S.B., & Bridgeman, B. (1986). Testing ESL student writers. In K.L. Greenberg, H.S. Wiener, & R.A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 126-152). New York: Longman.

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

Clausen, J. (1986). Warming up to proficiency: A project in process. *ADFL Bulletin, 18,* 34-37.

Cooper, C.R. (1977). Holistic evaluation of writing. In C.R. Cooper, & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-31). Urbana, IL: National Council of Teachers of English.

Dvorak, T. (1986). Writing in the foreign language. In B.H. Wing (Ed.), *Listening, reading, writing: Analysis and application* (Reports of the Northeast Conference on the Teaching of Foreign Languages, pp. 145-167). Middlebury, VT: Northeast Conference.

Elbow, P. (1986). Portfolio assessment as an alternative to proficiency testing. *Notes from the National Testing Network in Writing* (November), 3, 12.

Ellis, R. (1985). *Understanding second language acquisition.* Oxford, England: Oxford University Press.

Freed, B.F. (1987). Preliminary impressions of the effects of a proficiency-based language requirement. *Foreign Language Annals, 20,* 139-146.

Freedman, A., & Pringle, I. (1980). Writing in the college years: Some indices of growth. *College Composition and Communication, 31,* 311-324.

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

Hoffman, E.F., & James, D. (1986). Toward the integration of foreign language and literature teaching at all levels of the college curriculum. *ADFL Bulletin, 18,* 29-33.

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

Lloyd-Jones, R. (1982). Skepticism about test scores.

Notes from the National Testing Network in Writing (October), 3, 9.

Lowe, P., Jr. (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and, particularly, to Bachman and Savignon. *Modern Language Journal, 70,* 391-397.

Magnan, S.S. (1985). Teaching and testing proficiency in writing: Skills to transcend the second-language classroom. In A.C. Omaggio (Ed.), *Proficiency, curriculum, articulation: The ties that bind* (Reports of the Northeast Conference on the Teaching of Foreign Languages, pp. 109-136). Middlebury, VT: Northeast Conference.

Odell, L. (1981). Defining and assessing competence in writing. In C.R. Cooper (Ed.), *The nature and measurement of competency in English,* (pp. 95-138). Urbana, IL: National Council of Teachers of English.

Omaggio, A.C. (1983). *Proficiency-oriented classroom testing* (Language in Education series No. 52). Washington, DC: ERIC Clearinghouse on Languages and Linguistics. (ERIC Document Reproduction Service No. ED 233 589)

Osterholm, K.K. (1986). Writing in the native language. In B.H. Wing (Ed.), *Listening, reading, writing: Analysis and application* (Reports of the Northeast Conference on the Teaching of Foreign Languages, pp. 117-143). Middlebury, VT: Northeast Conference.

Purnell, R.B. (1982). A survey of the testing of writing proficiency in college: A progress report. *College Composition and Communication, 33,* 407-410.

Richards, J.C. (1984). The secret life of methods. *TESOL Quarterly, 18,* 7-23.

Rosengrant, S.F. (1987). Error patterns in written Russian. *Modern Language Journal, 71,* 138-146.

Ruth, L., & Murphy, S. (1984). Designing topics for writing assessment: Problems of meaning. *College Composition and Communication, 35,* 410-422.

Savignon, S. (1983). *Communicative competence: Theory and classroom practice.* New York: Addison Wesley.

Shuy, R.W. (1981). Toward a development theory of writing. In C.H. Frederiksen & J.F. Dominic (Eds.), *Writing: Process, development and communication* (pp. 119-132). Hillsdale, NJ: Erlbaum.

Stansfield, C. (1986). A history of the Test of Written English: The developmental year. *Language Testing, 3,* 224-234.

Van Oorsouw, J. (1986). *Assessing writing proficiency: Methods and assumptions.* Unpublished manuscript.

Wolcott, W. (1987). Writing instruction and assessment: The need for interplay between process and product. *College Composition and Communication, 38,* 40-47.

Zamel, V. (1976). Teaching composition in the ESL classroom: What we can learn from research in the teaching of English. *TESOL Quarterly, 12,* 67-76.

Zamel, V. (1982). Writing: The process of discovering meaning. *TESOL Quarterly, 16,* 195-209.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly, 19,* 79-101.