DOCUMENT RESUME

ED 172 523 FL 010 176

AUTHOR	Clark, John L. D., Ed.
TITLE	Direct Testing of Speaking Proficiency: Theory and Application.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	78
GRANT	G007701871
NOTE	261p.; Proceedings of a Conference conducted by Educational Testing Service in cooperation with the U.S. Interagency Language Round Table and the Georgetown University Round Table on Languages and Linguistics (Washington, D.C. March 1978)
AVAILABLE FROM	Educational Testing Service, Princeton, New Jersey 08541 (free)
EDRS PRICE	MF01/PC11 Plus Postage.
DESCRIPTORS	Achievement Tests; *Communicative Competence (Languages); Conference Reports; Higher Education; Language Fluency; Language Instruction; *Language Proficiency; *Language Skills; *Language Tests; Linguistic Competence; Secondary Education; Second Language Learning; Speech Communication; Test Construction; *Testing; Test Validity
IDENTIFIERS	*Oral Testing

ABSTRACT

The following papers are presented in the conference proceedings: (1) "Development and Current Use of the FSI Oral Interview Test," by H. Sollenberger; (2) "Interview Testing in Non-European Languages," by W. Lovelace; (3) "Measuring Second Language Speaking Ability in New Brunswick's Senior High Schools," by M. Albert; (4) "Using the FSI Interview as a Diagnostic Evaluation Instrument," by S. Graham; (5) "Direct Testing of Speaking Skills in a Criterion-Referenced Mode," by E. Franco; (6) "Oral Proficiency Testing in New Jersey Bilingual and English as a Second Language Teacher Certification," by R. Brown; (7) "Adaptation of the FSI Interview Scale for Secondary Schools and Colleges," by C. Reschke; (8) "Interview Techniques and Scoring Criteria at the Higher Proficiency Levels," by R. Jones; (9) "Testing Speaking Proficiency Through Functional Dialogues," by I. Roos-Wijgh; (10) "Scope and Limitations of Interview-Based Language Testing: Are We Asking Too Much of the Interview?" by R. Lado; (11) "Measuring Foreign Language Speaking Proficiency: A Study of Agreement Among Raters," by M. Adams; (12) "Independent Rating in Oral Proficiency Interviews," by J. Quinones; (13) "Third Rating of FSI Interviews," by P. Lowe, Jr.; (14) "Determining the Effect of Uncontrolled Sources of Error in a Direct Test of Oral Proficiency and the Capability of the Procedure to Detect Improvement Following Classroom Instruction," by K. Mullen; (15) "Reliability and Validity of Language Aspects Contributing to Oral Proficiency of Prospective Teachers of German," by R. Clifford; (16) "Interview Testing Research at Educational Testing Service," by J. Clark; (17) "Psychophysical Scaling of the Language Proficiency Interview, A Preliminary Report," by R. Vincent; and (18) "Setting Standards of Speaking Proficiency," by S. Livingston. (AMH)

DIRECT TESTING OF SPEAKING PROFICIENCY: THEORY AND APPLICATION

Proceedings of a Two-Day Conference Conducted by Educational Testing Service in Cooperation with the U.S. Interagency Language Round Table and the Georgetown University Round Table on Languages and Linguistics

John L. D. Clark, ed.

Educational Testing Service, Princeton, NJ
1978

The work reported herein was conducted under grant No. G00-77-01871 from the U.S. Office of Education, Department of Health, Education and Welfare, under the authority of Title VI, Section 602, NDEA.

CONTENTS

PREFACE

The eighteen papers presented in this volume form the written record of a conference on "Direct Testing of Speaking Proficiency: Theory and Application" held at Georgetown University on March 14-15, 1978. It was conducted by Educational Testing Service with the cooperation of the U.S. Interagency Language Round Table and the Georgetown University Round Table on Languages and Linguistics. Financial assistance for the conference and for publication of the proceedings was provided by the U.S. Office of Education under the authority of Title VI, Section 602, of the National Defense Education Act.

In the approximately twenty years since the initial development, by the Foreign Service Institute (FSI), U.S. Department of State, of the face-to-face language proficiency interviewing procedure and associated rating scale, use of this or related approaches to speaking proficiency measurement has become increasingly widespread, both within and outside the federal government. A partial list of current users of interview-based testing techniques includes, in addition to the FSI, ACTION/Peace Corps, Bank of Canada, Center for Applied Linguistics, Central Intelligence Agency, Chula Vista (Calif.) School District, Cornell University, Defense Language Institute, Educational Testing Service, Florida International University, Illinois Bilingual Service Center, Language Training Mission of Brigham Young University, Massachusetts Department of Education, National Security Agency, New Brunswick (Canada) Education Department, and New Jersey Department of Education.

In view of the increasing interest in and utilization of language testing techniques of the FSI type over the past several years, it was considered of possible value to bring together--through the medium of a formal conference directed exclusively to interview-based assessment techniques or other face-to-face testing procedures--major users of these techniques and other interested participants, both to review and discuss matters of common interest in direct speaking proficiency testing and to serve as a forum for the broader dissemination of information in this measurement area.

The conference presentations, reproduced here in their final printed form, deal with one or more of the five major topical areas: (1) practical applications of direct proficiency testing; (2) testing procedures, including performance rating scales and scoring techniques; (3) training and quality control of testers and raters; (4) validity and reliability of direct testing techniques; and (5) current and proposed research and development activities in direct proficiency testing.

The opening paper, by Howard E. Sollenberger--former director of the Foreign Service Institute, who was, as he puts it, "present at the creation" of the FSI interview--details the development of the interviewing and rating procedure and its past and current use by U.S. government agencies and discusses the scope of proper utilization of this technique. Appendix A of his paper reproduces the Absolute Language Proficiency Ratings that constitute the official rating scale for the FSI interview and that may be referred to as needed in the reading of other conference papers.

The next five papers provide a rather broad overview of the operational use of the FSI technique or adaptations of the technique in a wide variety of measurement applications in both government and private contexts. William Lovelace describes the use of FSI-type interviews to evaluate the host-country language proficiency of Peace Corps volunteers and discusses some of the special considerations involved in the use of English-medium training procedures to train interviewers and raters for testing in non-European languages.

Murielle Albert describes a province-wide system of interview-based language testing at the secondary school level for the New Brunswick (Canada) Education Department, and emphasizes both programmatic and individual-student benefits of a direct proficiency measurement approach.

The paper by Steven L. Graham gives an overview of the large-scale, intensive language program conducted at the Language Training Mission (Provo, Utah) and the procedures used by the LTM to initially train and subsequently monitor the performance of interview testers/raters; this is followed by a discussion of diagnostic checklists and other procedures used to provide feedback to individual examinees.

Robert B. Franco of the Defense Language Institute, Monterey, describes the recent (1976) revision of the DLI language assessment system, which emphasizes the use of criterion-referenced interviewing and role-playing situations to determine students' functional command of the spoken language.

The paper by Richard W. Brown summarizes the bilingual and English-second-language teacher certification requirements recently adopted by the state of New Jersey and describes the interview-based testing program through which the speaking proficiency of teachers and teacher candidates is measured for certification purposes.

The next four papers address a number of different aspects of the interviewing process and suggest certain changes in testing techniques, scoring procedures, or utilization of results, both to guard against possibly inappropriate applications of this measurement technique and to enhance the measurement value of the interview approach for situations in which its use can be recommended. The paper by Claus Reschke proposes an expansion of the interview rating scale to provide more detailed information on examinee performance, especially for use at the secondary school and early college levels, where the total range of performance is typically restricted to the lower (0 - 2+) portion of the total FSI scale.

At the other end of the proficiency spectrum, Randall L. Jones addresses the challenge of testing examinees at the higher (3+ - 5) score levels, and describes his experimentation with a variety of supplementary techniques, including low-frequency vocabulary testing, sentence repetition, and specified situational cues, to measure the sophisticated kinds of language behavior at issue in the upper regions of the FSI scale.

Ingrid F. Roos-Wijgh, of the Dutch National Institute for Educational Measurement (CITO), describes a test development project based on role-playing techniques that engage examinees in realistic dialogue situations for specified communicative purposes. This testing approach--although historically and operationally distinct from the FSI interview as it has developed in the United States--is of considerable relevance to the examiner/examinee "situation" that is often included as the final step in the interview process.

In his detailed and wide-ranging paper, Robert Lado undertakes an analysis of the nature and psychometric characteristics of interview-based testing procedures in comparison with alternative or supplementary approaches, including the use of objective tests to assess the listening comprehension aspects of an examinee's performance and discrete-item tests of grammar, vocabulary, and pronunciation when diagnostic information on these language aspects is desired--rather than or in addition to the more global appraisal of proficiency provided by the face-to-face interview.

The third series of papers, comprising the conference presentations of six authors, addresses in some detail the basic psychometric characteristics of the FSI-type interview (or adapted versions of the interview) as they are manifested in operational use of the interview technique in a variety of measurement contexts.

Marianne L. Adams presents the results of a detailed study of the interrater reliability of the interview process as carried out by French, German, and Spanish interviewers/raters at the FSI and cites a very high degree of scoring consistency for raters in these three language groups.

John Quiñones describes an adaptation of the interview scoring process that involves use by the raters of a graphic scoring scale that is seen to permit more fine-grained discrimination of examinee performance than is possible under the regular (categorical) rating system and to facilitate the combining and analysis of ratings assigned by two or more raters to a single examinee.

The paper by Pardee Lowe, Jr., summarizes a recent study in which "third raters" of proficiency interviews (i.e., any evaluators of a given interview other than those present at the original interview) were found--contrary to expectations--to be generally no more severe in their ratings than the original rating team, supporting the validity of "third ratings" as conceptually and operationally similar to those given during initial scoring.

Karen A. Mullen reports high interrater correlations for an FSI-type test using a modified rating procedure ("poor," "fair," "good," "above average," and "excellent" for each of the language aspects of listening, pronunciation, fluency, and grammar) and compares pre- and post-instruction interview scores for a group of undergraduate ESL students to similar scores on the Test of English as a Foreign Language (TOEFL). Results of this comparison are analyzed in terms of the nature and measurement purposes of the two types of instruments.

Ray T. Clifford describes the development of a modified interview rating scale synthesizing the FSI verbal descriptions with five other rating scales, and subsequently used in conjunction with a "Teacher Oral Proficiency" interview that was experimentally compared to a tape recording- and booklet-mediated speaking test (the MLA Cooperative Foreign Language Proficiency Test) with a group of prospective German teachers at the University of Minnesota. Results of this study provide comparative information on the interrater, intrarater, and test-retest reliabilities of the direct testing procedure vis-à-vis the more highly structured MLA test, as well as initial data on the convergent and discriminant validity of both testing procedures as applied to the diagnostic assessment of discrete aspects of language performance (grammatical control, vocabulary, pronunciation, and fluency).

The editor reports on several interview-based testing studies conducted at Educational Testing Service and discusses study results from the viewpoints of prediction of rater competence based on performance during rating training; scoring reliability of trained interviewers; relationship of interview scores to other measures of language competence; and duration of interview as related to the practicality, validity, and reliability of the interview process.

The two final papers address the use and interpretation of interview-based test results. Robert J. Vincent presents the results of a study in which experienced language teachers were asked to estimate the relative difficulty of training a beginning language student from "zero" to any given level on the FSI scale, or between any two pairs of levels on the scale. Perceived difficulty data of the type presented, together with empirically derived measures of language learning difficulty (such as total contact hours required to reach various FSI levels), is of considerable interest from a psycholinguistic standpoint and is also of practical value in promoting a more accurate and realistic conception on the part of language teachers and administrators regarding the difficulty and amount of training required to reach specified levels of language competence.

Samuel A. Livingston describes the operation and results of an empirically based study conducted in collaboration with the New Jersey Department of Education to assist the Department in the setting of "passing" standards for bilingual and ESL teacher candidates on the FSI-type interview used as part of the certification process in the state. In addition to presenting the results of the New Jersey study, the author discusses the standard-setting procedure on a more general basis and urges the use of this or a similar technique in any other important "decision-making" contexts involving the use of interview test results.

Numerous individuals and several different organizations contributed in a variety of ways to the initial planning and conduct of the conference and to the compilation of the conference proceedings.

I would first like to thank my friend and colleague of long standing, Mr. Protase E. Woodford--associate director of the International Office at Educational Testing Service and project director for the conference--for his initial perception of the appropriateness and usefulness of convening current users of the FSI interview technique and other face-to-face speaking proficiency measures to describe their own testing activities and to share information, insights, and mutual concerns with others involved in or interested in the potential applications of these measurement approaches. His continued interest and support at all stages of the conference are much appreciated and gratefully acknowledged here.

Both Mr. Woodford and I are in turn indebted to each of the other individuals and groups who helped make the conference a reality, most notably Mrs. Julia A. Petrov--Chief of the Research Program, International Studies Branch of the Division of International Education, DHEW/USOE and project officer for the conference--who, from the very beginning of discussions with her office and throughout the project period, fully supported the underlying rationale and purposes of the conference and provided valuable suggestions on its overall content, structure, and implementation.

The conference also benefited greatly in the early planning stages from correspondence and discussions with Dr. James R. Frith, dean of the School of Language Studies at the Foreign Service Institute and chairman of the Management Committee of the Interagency Language Round Table, and with Dr. Dorothy E. Waugh, chairman of the Testing Committee of the Interagency Language Round Table, and the other members of the Testing Committee. All of these contacts were of substantial value in identifying and seeking the representation at the conference of both government and nongovernment agencies known to be using the FSI interview technique or adaptations of it, and in identifying specific topics and potential presenters for the conference.

Dr. James E. Alatis, dean of the School of Languages and Linguistics at Georgetown University and chairman of the 1978 Georgetown University Round Table on Languages and Linguistics, lent his full support to the purposes of the conference and graciously arranged for the conference to be included as a presession component of the 1978 Georgetown University Round Table. He and his associate, Mrs. Carolyn Adger, made available highly suitable meeting facilities on the Georgetown campus and extended every personal and professional courtesy in the course of the conference sessions.

Valuable assistance in coordinating conference arrangements in the Washington area and in providing on-site administrative support during the two conference days were provided, respectively, by Dr. Tracy Gray and Ms. Ann Convery of the Center for Applied Linguistics.

Staff members at Educational Testing Service who made substantial contributions to the work of the conference or the preparation of the proceedings include my secretary, Mrs. Dolores Robinson, who was of inestimable assistance at all stages of the project; Mrs. Nancy Parr, who provided excellent editorial and proofreading support; and Vydec operators Mrs. Maryann Cochran and Mrs. Brenda Mahan, whose admirable diligence and indefatigability provided the camera-ready text of the proceedings.

A final acknowledgment and most heartfelt appreciation are expressed to all of the conference presenters, whose contributions are reproduced in this volume. If a slight semantic liberty can be permitted me, I would like to close these introductory paragraphs by stating that these individuals <u>were</u> the March 14-15 conference and <u>are</u> the present proceedings, which it has been my great pleasure and honor to assemble here.

J.L.D.C.

DEVELOPMENT AND CURRENT USE OF THE FSI ORAL INTERVIEW TEST

Howard E. Sollenberger

Director, Foreign Service Institute (retired)

DEVELOPMENT AND CURRENT USE OF THE FSI ORAL INTERVIEW TEST

Howard E. Sollenberger

I address you today, not as a specialist in foreign language testing or as a linguist, but rather as an administrative philosopher and historian. Since I no longer administer, I can perhaps be permitted to give you some history of the development of the foreign language oral interview tests of the Foreign Service Institute (FSI) and to philosophize on the subject of this conference, "Direct Testing of Speaking Proficiency: Theory and Application."

I hope I am not presumptuous in assuming that a brief historical case study of the circumstances under which direct interview testing was first attempted on any significant scale, and how it developed into a system used throughout the federal government, would be helpful as background for our deliberations. Certainly we will want to examine both the advantages, and the implications, of putting theory into practice, in institutionalizing systems by which we attempt to measure and differentiate human performance.

To paraphrase Dean Acheson, you might say that I was "present at the creation" or, perhaps more accurately, at the incubation of the oral interview testing system developed at the FSI. While it may now be rather dim in our memories, we were in a period of "cold war" intensification in the early 1950s. It had wide and significant ramifications in our public life, and even in education. By the late 1950s it would, among other things, generate the National Defense Education Act, which was to support the upgrading of science, mathematics, and foreign area and language studies in American education. Meanwhile, with the impetus of the Korean War and the experience of having been unprepared for the global war a decade earlier, the Civil Service Commission in 1952 was directed, under the National Mobilization and Manpower Act, to inventory and develop a register of persons in government who had skills, background, and experience in various foreign areas and languages.

Following normal bureaucratic procedures, the Civil Service Commission created an interagency committee to study the problem and recommend procedures. At early meetings it became apparent that, if an inventory were to serve any useful purpose, some means of defining and differentiating levels of foreign language proficiency and area expertise would be necessary. The old labels of fair/good/fluent/bilingual were obviously inadequate.

Dr. Henry Lee Smith (then dean of the FSI Language School), the State Department's representative of the interagency committee, pressed for a system and the development of criteria that would differentiate testable levels between "no knowledge" of a given foreign language and "total mastery." He was promptly named to head a subcommittee to prepare definitions and so-called working papers. As Dr. Smith's alternate on the committee, I became involved as a coconspirator in trying to get the federal government to realistically face personnel deficiencies in area expertise and foreign language skills.

As it developed, there was not only difference of opinion, but also opposition to the concept. There was concern in certain agencies that through the proposed survey and the establishment of a national register, the Civil Service Commission would further interfere in the personal fiefdoms of the various agencies. There was also fear that testing based on new absolute standards would prove embarrassing to many employees who had claimed "fluency" in a foreign language on their applications for employment. To make a long story short, a compromise was reached that provided for each agency to conduct its own survey using definitions and criteria established by the committee. Testing would be optional.

There were five different factors considered in defining and differentiating levels of area expertise: systematic area training (A), basic social science training (S), professional experience in an area (PA), professional experience related to an area (PE), and residence in an area (AR). Three to five differentiated levels were defined under each factor.

Under the language proficiency section, symbolized by the letter L, six differentiated levels were defined. To avoid complicating the task, no effort was made to separate the components of language proficiency, which were generally considered to be comprehension of oral production, speaking proficiency, reading proficiency and comprehension, and writing. At the base of the scale, L-1 was defined as "no proficiency in either reading or speaking a foreign language."

The upper end of the scale, L-6, was defined as "sufficient proficiency in speaking, reading and writing to negotiate oral and written agreements and to thoroughly understand the press, popular and classical literature and official documents." It was noted that "this category is reserved for bilingual or native speakers of the language."

It was proposed that category L-4 be considered as the minimum proficiency level for inventory purposes. This was defined as "sufficient proficiency in speaking a language to conduct ordinary routine business conversations and to read general non-technical material." It was noted that "this level of proficiency might normally be acquired by 9 to 12 months of intensive language training or the equivalent in part-time study, depending on the difficulty of the language."

Bureaucratic foot-dragging, a change in the administration, and winding down of the Korean War resulted in the whole project being shelved.

However, at the FSI, enough interest had been generated in the potential usefulness of this approach to stimulate further refinement of the scale and to experiment with structured oral interview testing of students.

The second impetus came in 1955, when Loy Henderson, then Deputy Undersecretary of State, decided to conduct a survey of foreign language skills in the Foreign Service. Up to that time there had never been an inventory of language skills in the Foreign Service. Mr. Henderson was motivated by a conviction that post-war diplomacy would increasingly require face-to-face communication with people around the world as well as between government representatives and diplomats. In spite of some opposition within the Foreign Service, Mr. Henderson insisted that the survey be followed by testing. He also intended to tie promotions to tested foreign language proficiency. This was serious business in the highly competitive Foreign Service. It was also serious business for the FSI and those who would design and conduct the tests.

Testing of the 1952 definitions of L-1 through L-6 on some 200 officers showed them to be inadequate for the purpose of a self-appraisal survey of the Foreign Service. It became apparent that speaking and reading proficiencies would have to be separately determined. From this emerged the L and R scales, with the speaking (oral production) scale (L) differentiated from 1 to 6, and reading facility (R) differentiated from 1 to 5.

With this instrument a self-appraisal survey was conducted in the Foreign Service. It revealed that less than half of the 4,041 regular, reserve, and staff officers surveyed had a "useful to the service" proficiency in French, German, or Spanish. (These three languages, along with English, were considered the "world languages" of diplomacy.) "Useful" was then defined as "sufficient control of the structure of a language, and adequate vocabulary, to handle routine representation requirements and professional discussions within one or more special fields, and--with the exception of such languages as Chinese, Japanese, Arabic, etc.--the ability to read non-technical news or technical writing in a special field." This was the L-4, R-3 level as defined in the self-appraisal scales.

These findings led to a new language policy, announced by the Secretary of State on November 2, 1956. This policy was based on the premise that foreign language skills are vital in the conduct of foreign affairs. Therefore, "each officer [would] be encouraged to acquire a 'useful' knowledge of two (2) foreign languages, as well as sufficient command of the language of each post of assignment to be able to use greetings, ordinary social expressions and numbers; to ask simple questions and give simple directions; and to recognize proper names, street signs and office and shop designations." It further stated: "Evidence of achievement will be verified by tests administered by the Foreign Service Institute."

Having been committed to testing, FSI was under pressure to develop reliable test procedures. As Claudia P. Wilds pointed out in her paper "The Oral Interview Test," published in 1975 by the Center for Applied Linguistics in Testing Language Proficiency: "Both the scope and the restrictions of the testing situation provided problems and requirements" previously unknown in language testing.

In the course of developing and refining oral interview test procedures, Professor John B. Carroll, then of Harvard, was consulted. This led to a revision of the differentiated levels of proficiency and the redesignation of the symbols and levels. The symbol L was changed to S to identify the scale for speaking proficiency. R remained the symbol for the reading scale. Each scale was differentiated into six levels, numbered from 0 to 5.

Since this provided, for the first time, officially approved performance and criterion-based definitions that testers, instructors, and administrators found useful, the system rapidly became institutionalized and the S and R symbols became part of the jargon.

Not surprisingly, problems began to emerge. Officers being tested complained that different testing teams applied different standards, particularly in testing different languages. For example, it was commonly believed--and with some justification--than an S-3 rating was much tougher to get in French than in the so-called hard or esoteric languages. It was also rumored that students tested by their own instructors seemed to fare better than those who simply came in for tests. Testers seemed to be more critical in judging the performance of those whom they did not know through a teacher-student relationship. In some cases, the rank and age of the officers were seen to influence the rating. Informally there developed what became known as the "compassionate" S-3 rating. There was also evidence that some testers seemed to be unduly influenced by the personalities and cooperativeness of persons being tested.

With mandatory testing of Foreign Service officers announced in 1957, and with assignments and promotions to be influenced by the results, these problems had to be solved. An independent testing unit was established in July 1958, with Frank A. Rice as head of the unit and Claudia Wilds as his assistant. It was through the collaboration of these two people that a significant breakthrough came in standardizing oral testing procedures. A checklist was developed that contained five "factors": accent, grammar, vocabulary, fluency, and comprehension. Considerable work went into selecting these factors. The criterion was that they should be of a sufficiently general nature that they would apply equally well to all languages. Each factor was subdivided as a six-point descriptive scale, with "polar" terms X (extremely poor or inadequate) and Y (extremely good, accurate, or complete).

As Frank Rice pointed out in an article entitled "The Foreign Service Institute Tests Language Proficiencies" (Linguistic Reporter, May 1959): "The original purpose of the Check List was to help counterbalance the inherent subjectivity of the testing procedure by providing agreement about what aspects of the performance were to be observed, a control on the attention of the observers, and a system of notation that would make judgments of different observers more nearly comparable.

"There is no doubt that the Check List accomplished its original purpose. This was expected. What was quite unexpected was what emerged from statistical analysis. This provided basic evidence of a high degree of consistency in the subjective judgments of the examiners. The instrument could thus serve not only as a useful record, but also as a highly accurate predictor."

It also provided a means for training testers. Claudia Wilds, who was appointed head of the testing unit in 1963, subsequently developed a weighted scoring system for the checklist. Among other things, this provided a means for occasional verifications of the checklist profiles and seemed to keep examiners in all languages reasonably in line with each other.

Further evidence of the success of this system was the sharp drop-off of complaints from persons being tested, and general acceptance of the results even for critical personnel decisions. Also, use of the rating scale and test results began to spread. With some modifications, the CIA developed a similar system, and the United States Information Agency and the Agency for International Development joined with the Department of State in using the FSI-developed standards and testing facilities.

Even the Congress used them, demanding reports based on FSI standards to show progress toward compliance with a legislative mandate that the Department of State "designate every Foreign Service officer position in a foreign country whose incumbent should have a useful knowledge of a language or dialect common to such country [and that] each position so designated... be filled only by an incumbent having such knowledge" (Sec. 578 Foreign Service Act of 1946).

With the spreading use, in the 1960s, of the proficiency rating scale to other agencies, including the Defense Language Institute and the Peace Corps, it became apparent that the definitions should be further revised and standardized among agencies. Representatives of the FSI, the CIA, the Defense Language Institute, and the Civil Service Commission met in 1968 and developed a unified version of the definitions. These definitions are essentially the ones used today, and are shown as Appendix A of this paper.

Now, twenty-five years after the inception of a criterion-referenced rating scale, it has been incorporated into the federal personnel manual for use throughout the U.S. government, and it has been adopted by the Supreme Headquarters of the Allied Powers in Europe. Educational Testing Service has joined the ranks of users, and increasing interest has been shown in academic circles--an interest that promises impact and contributions in the future.

At the beginning of this paper, I stated my hope that we would examine the limitations and implications of applying theory to practice in the direct testing of speaking proficiency. As I have observed this in the government, it has become apparent to me that one of the principal limitations is the inability of this system to make meaningful judgments or to measure the most significant objective of human speech--effective communication. By this I mean the effectiveness or lack thereof of an individual in listening to and fully understanding what he hears through the static of cultural differences and the peculiarity of personality, and the ability to communicate fully with another person of a different culture in such a way as to achieve understanding and cooperation.

I have observed more than a few cases where I cringed at the thought that an individual would represent the United States overseas, even though he had been given a high S-4, R-4 language proficiency rating by our tests. The person's so-called language proficiency, while it may have been quite accurate in terms of technical skill, did not mean effectiveness in communication. In some cases, it may have enabled the person to misrepresent or foul up more effectively. This is to say that you can be a fool in any language or that you can put your foot in your mouth in any language. Nor does the fact of technical ability to use a foreign language without noticeable accent or grammatical errors mean that the person has something worth saying. I'm sure we all know people who talk nonsense fluently.

On the other hand, I know people who butcher the language, whose accents are atrocious, and whose vocabularies are limited. For these reasons we give them low proficiency ratings. Yet, for some reason, some of them are effective communicators.

You may rightly say that the tests we have developed do not measure this dimension of effective communication. Still, I know a number of administrators and even some linguists who do not understand the imolication of this difference.

I have also observed, in the application of these testing procedures in training situations, a tendency to train for success on the test score, or to the standards of the test, rather than for broad effectiveness in communication. It becomes more important to the teacher and the student that they achieve the S-3 level, rather than that they be effective communicators. These are not necessarily mutually exclusive objectives, but there are times when this is forgotten.

I am not saying that these limitations, which deal with the use of measurement devices we create, should cause us to abandon our efforts to perfect and use such systems. It is, however, my conviction that these and other limitations must be recognized and that we have a continuing obligation to make these limitations known to end users. In this we are no different from the scientist who makes a discovery that can, if properly used, be of benefit to human kind but that can also be misused. I hope this conference will not ignore these responsibilities.

Appendix A

Absolute Language Proficiency Ratings[1]

The rating scales described below have been developed by the Foreign Service Institute to provide a meaningful method of characterizing the language skills of foreign service personnel of the Department of State and of other Government agencies. Unlike academic grades, whict measure achievement in mastering the content of a prescribed course, the S-rating for speaking proficiency and the R-rating for reading proficiency are based on the absolute criterion of the command of an educated native speaker of the language.

The definition of each proficiency level has been worded so as to be applicable to every language; obviously the amount of time and training required to reach a certain level will vary widely from language to language, as will the specific linguistic features. Nevertheless, a person with S-3's in both French and Chinese, for example, should have approximately equal linguistic competence in the two languages.

The scales are intended to apply principally to government personnel engaged in international affairs, especially of a diplomatic, political, economic and cultural nature. For this reasor heavy stress is laid at the upper levels on accuracy of structure and precision of vocabulary sufficient to be both acceptable and effective in dealing with the educated citizen of the foreign country.

As currently used, all the ratings except the S-5 and R-5 may be modified by a plus (+), indicating that proficiency substantially exceeds the minimum requirements for the level involved but falls short of those for the next higher level.

[1] FSI Circular, November 1968.

Definitions of Absolute Ratings

Elementary Proficiency

S-1 Able to satisfy routine travel needs and minimum courtesy requirements. Can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak his language; while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at the S-1 level should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases, and tell time.

R-1 Able to read some personal and place names, street signs, office and shop designations, numbers, and isolated words and phrases. Can recognize all the letters in the printed version of an alphabetic system and high-frequency elements of a syllabary or a character system.

Limited Working Proficiency

S-2 Able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e. topics which require no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.

R-2 Able to read simple prose, in a form equivalent to typescript or printing, on subjects within a familiar context. With extensive use of a dictionary can get the general sense of routine business letters, international news items, or articles in technical fields within his competence.

Minimum Professional Proficiency

S-3 Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Can discuss particular interests and special fields of competence with reasonable ease; comprehension is quite complete for a normal rate of speech; vocabulary is broad enough that he rarely has to grope for a word; accent may be obviously foreign; control of grammar good; errors never interfere with understanding and rarely disturb the native speaker.

R-3 Able to read standard newspaper items addressed to the general reader, routine correspondence, reports and technical material in his special field. Can grasp the essentials of articles of the above types without using a dictionary; for accurate understanding moderately frequent use of a dictionary is required. Has occasional difficulty with unusually complex structures and low-frequency idioms.

Full Professional Proficiency

S-4 Able to use the language fluently and accurately on all levels normally pertinent to professional needs. Can understand and participate in any conversation within the range of his experience with a high degree of fluency and precision of vocabulary; would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations; errors of pronunciation and grammar quite rare; can handle informal interpreting from and into the language.

R-4 Able to read all styles and forms of the language pertinent to professional needs. With occasional use of a dictionary can read moderately difficult prose readily in any area directed to the general reader, and all material in his special field including official and professional documents and correspondence; can read reasonably legible handwriting without difficulty.

Native or Bilingual Proficiency

S-5 Speaking proficiency equivalent to that of an educated native speaker. Has complete fluency in the language such that his speech on all levels is fully accepted by educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialisms, and pertinent cultural references.

R-5 Reading proficiency equivalent to that of an educated native. Can read extremely difficult and abstract prose, as well as highly colloquial writings and the classic literary forms of the language. With varying degrees of difficulty can read all normal kinds of handwritten documents.

INTERVIEW TESTING IN NON-EUROPEAN LANGUAGES

William Lovelace

ACTION/Peace Corps

INTERVIEW TESTING IN NON-EUROPEAN LANGUAGES

William Lovelace

One of the most important aspects of overseas service as a Peace Corps volunteer is the ability to speak a foreign language or languages. Indeed, two of the three Peace Corps goals relate to an improved understanding between Americans and peoples of the world. Training and evaluating our volunteers in these languages has been a unique challenge to the agency, given the large number of languages volunteers are asked to learn (at least twenty in Africa alone) and the fact that these languages are often little-known and rarely studied.

A further complication, particularly in Africa, is that the volunteers must be trained and tested in both the official (European) language and the local language. The European language is almost always a Romance language. I say almost always since English is the official language of nine African countries as well as Belize, the Eastern Caribbean, Jamaica, and several areas of the Pacific. Even in these Anglophone countries, however, English is not always the language most appropriate for village-level communication, and proficiency in a local language becomes necesssary if the volunteer is to be effective.

Evaluating the proficiency of our volunteers in the various languages of the world is a challenging assignment, and analyzing the language levels in Anglophone countries has proven to be particularly difficult. The history of our language evaluations has, to some degree, resulted from our training formats.

During the early years of the Peace Corps, the majority of the training programs took place at university campuses. This classroom instruction was compatible with the FSI interview format, and we used FSI testers to interview volunteers in French, Spanish, and Portuguese. As we shifted the training emphasis to in-country, we had an increased need to test in the many national languages our volunteers learn. This meant that we could no longer use imported FSI testers; we needed to rely on host-country testers to interview volunteers in these languages. Our initial contract with Educational Testing Service (ETS), therefore, called for not only interviews of language students but also certification of testers. However, the certification of testers in "exotic" languages that the certifiers did not speak became a definite complication. For those countries in Latin America and Africa where Romance languages are spoken, these languages were used as certification vehicles for the local languages. These Romance languages, however, are not appropriate to Asia and to Anglophone countries in the rest of the world. We therefore had recourse to certification through English for these situations.

The use of a European language as a test medium raises questions, some of which I will discuss. No matter what theoretical or philosophical constraints we may face in this testing procedure, we feel we must evaluate all our volunteers. This is in part due to fiscal responsibility. We spend a large part of our training budget on language, and we are held responsible for tracking the results of these expenditures. This money is spent to train in many world languages. At one time we were not equipped to train and test in the non-Romance languages. We realized, however, that testing and training are absolutely essential for all volunteers if we are to honor the commitment contained in the Peace Corps goals I mentioned earlier. In surveys taken of the volunteers, we are reminded of this need.

The annual survey of volunteers has recently been published, and it contains data that are specifically relevant to our language training. The study shows a strong correlation between job satisfaction/ psychological well-being and an ability to speak the local language. The survey also shows a direct connection between satisfied volunteers and training programs incorporating home stays with host-country families (with a high priority on local language). Further, the survey shows that 55 percent of the respondents throughout the Peace Corps use a non-English host-country language at least half the time in their work. Also, as a group, the volunteers who are least satisified with their language training serve in Anglophone countries. It is therefore in these countries that we perhaps have the most to accomplish in training and evaluation.

But, it is also in these countries that we face the challenge of certification of testers in English. In our original agreement with ETS, if someone were certified in French, Spanish, or Portuguese, that person was also certified to test in one or more local languages. We decided to maintain this practice and to use a certification kit of listening tapes and ETS visits to certify testers in the Anglophone countries. Some of the following points of discussion relate to our certifying testers in European languages, but there are some ideas specific to certification in English that I wish to stress.

There is no doubt that the certification by ETS of host-country testers adds an element of "status" and a sort of professional recognition to those people working for the Peace Corps overseas. It must be admitted that Peace Corps employment is not always seen as representing any sort of professional standing, and our working relationship with an institution such as ETS lends credibility to our language program. In Africa, without certification through English, we would be unable to have this recognition in non-Francophone countries. The use of ETS certification helps assure that we have a standardized and widely recognized "shorthand" for language testing throughout Africa and across linguistic lines. This in turn enables the volunteers in Anglophone countries to enjoy the same advantage of Francophone volunteers: a record of their language proficiency can be kept on file at ETS in Princeton. Admittedly, an official 2+ in Krio or siSwati may have less "clout" and be less valuable for graduate credit than a similar score in French or Spanish, but this record can represent a tangible acquisition after two years of volunteer service.

Testing volunteers in the host language also adds a professional note to our in-country language programs. The volunteers are more likely to apply themselves in their language studies if they know they are being "rated." There is often a spirit of competition and pride in the language programs that would not exist without a record of progress. This situation is true for all volunteers, who must be able to deal with local and village situations, but it is particularly helpful to volunteers serving in countries where one can coast or "get by" in English.

The use of English for certification of testers has caused some concern among those of us in the Peace Corps working in language programs. Perhaps the most obvious issue is that this process requires that the candidate have a rather sophisticated level of English; he or she must be able to successfully rate the test tapes. English is widely spoken in many overseas countries and, as I mentioned, is an official language in large parts of the Peace Corps. However, limiting the group of possible testers to those demonstrating an ability to analyze English does place a severe constraint on the pool of applicants. There is also the fact that in certifying someone to test in English (or a Romance language) we have no guarantee that this demonstrated ability to analyze French or English can be transferred to the candidate's non-European native language. We must use this inferred ability to shift analytical skills as a base in our use of ETS certification since, with the exception of Latin America and parts of the Pacific, our tester-candidates are not native speakers of a European language. It is unrealistic to develop tester certification in the many languages volunteers work in, including such national languages as Thai, Farsi, and kiSwahili.

In Africa, the use of English for certification also spotlights the fact that Americans are certifying Africans in a language that differs somewhat in the various parts of the world. The English spoken throughout Africa can certainly be evaluated against standard norms of "correct" English, but there is a wide range of accents and vocabulary among Africans who live thousands of miles apart. The use of English also brings out the issue that we are certifying testers in a language they will never be asked to test in. We will probably never request a host-country tester to evaluate a volunteer's English level.

A further assumption we have made in certifying in English is that the person who "passes" the English certification is able to go through the same thought process in his or her native language. In Anglophone Africa we are often dealing with an indigenous language that the native speaker has not studied as an academic subject; a language that may be neither written nor read.

The nature of many of the African languages has raised the concern that these somewhat exotic languages do not necessarily lend themselves to an FSI-type interview analysis. Some of these languages are little known or studied other than in linguistic or perhaps missionary circles, and there is probably little information available as to the structure, patterns, and elements that would constitute a 2+ in Mende. Our experience shows that the tester is usually so taken by the volunteer's ability (and desire) to speak a language not often studied by outsiders that the ratings depend almost entirely on fluency, nonverbal social cues inherent in the language, and the use of proverbs or vignettes that reflect the history or philosophy of the culture.

There is, finally, the concern that an FSI interview in a local language is not related to the everyday use of the language by the volunteer. These languages are normally used in job-specific settings. They would not be used in high-level or official contacts and would rarely have the kind of direct question/answer format of the traditional interview.

Having outlined reasons why we feel we must evaluate our volunteers' proficiency in foreign and sometimes exotic languages, and having discussed some of the questions raised by certifying host-country testers in European languages (and especially in English), I unfortunately have little to say about what we are doing to change things. I believe we should give more thought to situational testing, which would be more closely related to a volunteer's use of the language. To do this, we would have to change our test format. We would also have to develop criteria for rating someone's ability to perform a set exercise in the foreign language and, if possible, equate that performance to a scale that would have outside recognition, such as an FSI level. This is a challenge facing the Peace Corps, one which the new administration of the agency may choose to face in the near future.

MEASURING SECOND LANGUAGE
SPEAKING ABILITY IN
NEW BRUNSWICK'S SENIOR HIGH SCHOOLS

Murielle Albert
Education Department
(New Brunswick, Canada)

MEASURING SECOND LANGUAGE SPEAKING ABILITY IN NEW BRUNSWICK'S SENIOR HIGH SCHOOLS

Murielle Albert

Introduction

English is not the sole language spoken in the province of New Brunswick, Canada. About 34 percent of New Brunswick's population, which is now close to 700,000, are French-speaking. So, for many of these New Brunswickers, Enjlish is a second language--that is, a language necessary for certain official, social, commercial, or educational activities within their own province and country.

On July 1, 1977, New Brunswick officially became a bilingual province. Therefore, English is a top requirement of those seeking good jobs within the province and is the language in which most of the business affairs of the more prestigious and more highly paid jobs are conducted in other provinces of Canada.

Background

English as a second language has always been taught in New Brunswick's schools. Students generally have the opportunity of learning English for a minimum of six years tc a maximum of nine years before they leave high school.

Unfortunately, until six or seven years ago, students leaving high school with six to nine years of English could hardly communicate in the target language among themselves and even less with English-speaking people. Too much stress had been placed on the reading and writing skills and not enough on the listening and speaking skills. As a result, the Department of Education decided to introduce new programs in New Brunswick schools stressing oral proficiency, as summarized in Appendix A. (New programs were also introduced for French as a second language.)

I was teaching at the high school level at the time and was asked by my superintendent to pilot one of these new courses taught by the aural-oral approach. Having accepted, I spent a few summers studying this new approach and became what we call a language model.

The audiolingual objectives were to teach the student to comprehend the language when spoken at normal speed; to speak with "near-native pronunciation and intonation"; to read and write "with minimal recourse to bil.ngual dictionaries"; and to "understand" the people, their culture, and their heritage.

Truly, we, the foreign language teachers, had come a long way. Once absorbed in what we were going to do in the classroom, we were now more interested in what we could make possible for students to do there to develop their communicative competence as well as an awareness of cultural and ethnic differences. We were therefore charged to provide learning activities in which students used the language and to assure that students were having the best language experience possible, commensurate with their abilities, interests, and age levels.

The Department and the teachers were very excited about the new program, which proved to be the answer to their idea of learning a second language. To stress the importance of oral competency in the minds of the students and teachers alike, the Department decided to evaluate the spoken English as a second language (EASL) or French as a second language (FASL) of New Brunswick's high school population. Previously, the evaluation of English as a second language had been a written evaluation that basically tested the reading and writing aspects of the language. The listening and speaking skills had never been evaluated as such.

How was this to be done? No oral testing program existed in any of the other Canadian provinces. So, the only program to be tried was the interview procedure developed by the Foreign Service Institute and administered by Educational Testing Service for the Peace Corps and other programs.

Training

The purpose of training New Brunswick second language teachers to do the interviewing was to ensure that New Brunswick teachers would have as much involvement with the program as possible and, perhaps most important, as a result of the training and practice to contribute to their professional development as second language teachers. It was assumed that teachers who had such a close involvement with the program would be supportive of the program and that maximum cooperation would result.

To train teachers as classified interviewers for the province, practice tapes as well as testing tapes had to be made available. The voices on the tapes had to be those of our students, interviewed by classified interviewers from ETS. And that is how I came to have the pleasure of meeting and working with Russ Webster and Woody Woodford.

Russ and Woody came to my school, in Caraquet, in the spring of 1974 to interview and tape sixty students. I don't know who enjoyed those sessions more, the interviewees or the interviewers. The students would come out of the interviews beaming with excitement. Most of them would rush to me and tell me how friendly the two interviewers were--how they had made them laugh and actually forget they were speaking English. Even the 0 level student felt very much at ease and thought he had performed well. The experience proved to be very successful and I, personally, was excited about the whole program.

To date, there have been four training sessions. The first two were part of the initial contract with Educational Testing Service; the others were added in 1977 and 1978 due to the increased demand for the interviews.

A summary of the results of these training sessions follows.

Session No. 1

(According to the contract with ETS, this session would train twenty New Brunswick teachers to administer interviews and in turn to train other second language teachers in the province.)

No. Enrolled		No. Qualified		No. of Trainers	
FASL	EASL	FASL	EASL	FASL	EASL
10	10	2 + 4 (6)	3 + 7 (10)	4	7

Session No. 2

No. Enrolled		No. Qualified		No. of Trainers	
FASL	EASL	FASL	EASL	FASL	EASL
19	27	3 + 7 (10)	3 + 11 (14)	7	11

Session No. 3

No. Enrolled		No. Qualified	
FASL	EASL	FASL	EASL
28	17	9	11

Session No. 4

No. Enrolled		No. Qualified	
FASL	EASL	FASL	EASL
10	16	10	15

In addition, ten individuals who did not qualify at the time of the training resubmitted the test tapes to ETS.

No. Resubmitting Tapes		No. Qualified	
FASL	EASL	FASL	EASL
8	2	7	2

We now have forty qualified interviewers for French as a second language and fifty qualified interviewers for English as a second language. Included in these totals are the eleven French-as-a-second-language trainers and eighteen English-as-a-second-language trainers. I must add here that these teachers were all invited to participate. It wasn't thrown open to all second language teachers.

The training sessions lasted two to three days. During that period of time the teachers, guided by resource people from ETS, discussed the technical and linguistic aspects of the language proficiency interview, the assignment of interview ratings/discussion of student performance, and the numerical rating procedure. Then the recordings of the practice interviews were played and the teachers scored them to the best of their ability. This was followed by a discussion of the scoring of the abovementioned interviews as they are described in the manual.

The next step was the formation of groups of about six to eight teachers for the live interviews. Enough pupils were brought in so every teacher had the opportunity to interview one pupil. While the interviews were being recorded the observing teachers and the trainers scored the performance. After an interview, the raters discussed both the interview techniques and the scoring. By the time the live interviews were over, most groups were able to reach basic agreement on methods and standards, thereby ensuring a reasonable degree of uniformity.

The final step in the training session was scoring the test tapes. Each teacher was given ten tapes to score independently, with the help of the manual. These test tapes were sent to ETS to be evaluated. Whether a person qualified depended primarily on one's success with the test tapes and one's ability to interview effectively during the live performance.

As the teachers weren't too sure what the workshops were all about, many were apprehensive and didn't perform to the best of their abilities. (To be frank, it isn't a normal situation.) Moreover, they had to perform in the target language, and many teachers felt that their spoken English was a bit rusty. As one teacher remarked, "If I had been tested beforehand and knew my level of proficiency, I'd have more confidence." Many told me afterwards that the only English they spoke was in the classroom so they lacked vocabulary when it came to testing higher-level students.

I had the opportunity to meet those teachers within the next year, and most of them told me how valuable the experience had been for them. As they were classified interviewers, they had a good idea what the spoken proficiency of their students was, and their goal was to raise the level of proficiency. Many of these teachers succeeded in organizing some sort of oral testing program in their schools. Others couldn't organize any because they felt it was too time-consuming, especially in the larger high schools. But as the interviewers discussed the program with the other teachers, an awareness was born and oral production became the primary skill to be stressed in our schools.

After the first oral interview evaluation, in the spring of 1976, teachers realized that in the time that had elapsed between training and the first day of interviewing, many of the skills had become nebulous. For instance, many teachers would spend the entire day interviewing and then most of the night relistening to the tapes to be sure of their evaluation. It was suggested that time be allotted to recalibrate the interviewers between the training sessions and the actual interviews.

One such recalibration session was held in January of this year. Once again the teachers were invited to participate, and of the fifty qualified interviewers twenty-nine participated. (Some were ill; others were snowbound.) So many teachers responding so well to the call could only mean that they were really concerned and felt the need to be recalibrated. (I think I should mention here that some teachers had to drive close to 400 miles round trip.)

The recalibration session was similar to the training session. It was a two-day period designed to permit the interviewers to fully review the interview techniques. Once again the teachers, guided by resource people from ETS and local trainers, reviewed the technical and linguistic aspects of the language proficiency interview, did live interviews, and scored new test tapes to be evaluated by ETS. The session was also profitable as it was the first time all the qualified teachers were working together and the exchange of ideas was invaluable. At the end of the two days, the teachers felt better prepared to begin the spring testing program.

Scheduling

The oral interview evaluation is scheduled for the spring of each year, from March to May, inclusive. The high schools are invited to participate; it is not compulsory. So far, we've had two testing sessions. In the spring of 1976, out of the 68 high schools in the province, 7 did not request service and interviews were completed in approximately 51. A total of 2,466 students were tested: 1,386 EASL and 1,080 FASL. In the spring of 1977, of the 68 high schools, 7 did not request service and interviews were completed in approximately 50; 3,417 students were tested (1,927 EASL and 1,490 FASL). We foresee a few more schools for this spring.

The schools taking advantage of the service are contacted by the interviewers assigned to them; arrangements are made regarding, for example, exact dates of the testing, available space needed, and materials to be used (tapes and tape recorders).

Teachers also have to be given sufficient lead time to reacquaint themselves with the interview technique. For this purpose, each teacher receives a box of the practice tapes plus the manual containing the discussion of the practice interviews and the description of the language interview program. Some teachers meet together to play the previous interviews, to score them, and to discuss and agree upon the general method of conducting the interview. The others review by themselves. This is done to ensure a reasonable degree of uniformity.

In order to release qualified teachers to do the interviews, the Department defrays the cost of substitute teachers. Sufficient time is also needed for the teachers to prepare work for their students and substitute teachers. Most school districts do not want their teachers away for more than five days. It is up to each interviewer to decide whether he or she will evaluate for five consecutive days or intersperse the days.

Instructions to Principals and Students

In schools where there are classified interviewers, students meet collectively before the interviews begin and the interviewer answers any questions and tries to put the pupils at ease. Principals are also made aware of the needs of the interviewers prior to their arrival in the schools. Requirements include rooms away from traffic noise, interviewing cards filled out beforehand, and good sound equipment.

Taping Oral Interviews

Each interview is taped individually with tapes provided by the Department. (Tape recorders are also provided if the schools and teachers do not have any.) After the interviews, the tapes are kept at the Department like any other departmental grades. They are used for continuing inservice programs, and can be used by prospective employers or universities if the students concerned are willing to release them.

Our experience has been that the length of an interview depends on the student's ability to communicate. A poor student might be interviewed for twelve or fifteen minutes, while a good student might be interviewed for up to twenty-five minutes. Some teachers take five or ten minutes after each interview to finalize their ratings; others give a tentative score and take the tapes home to listen to again before giving their final ratings.

The candidates tested were the students in grades 11 and 12 in all the courses: academic, industrial, home economics, and commercial. Each teacher interviewed about fifteen students a day. The smaller high schools completed the interviewing in a week or less; it took somewhat longer in the larger high schools. The interviewers made sure that everyone was tested as they did not want any students to feel left out of the activity.

Student and Teacher Reaction

The results of the first two evaluation sessions were very positive. As with all exam results, some students were pleased and others were not.

Many who were not pleased with their performance and who were returning to school had a goal to work for. If they scored a 1 their aim was to reach 1+ or 2; if they scored a 2+ their aim was to reach 3 or 3+.

The results were also a revelation for some teachers. They realized that the oral production of their classes was either good or bad and decided to do something about the mediocre performance of some of their students.

In many of the large high schools, the department heads and the teachers concerned made detailed studies of the results. If, for example, 35 percent of the students had scored 1+, the objective of the English department for the following year was to try to raise the level to 2 or 2+.

As these tests are province-wide, each school knows where it stands on a provincial basis. So another incentive for schools is to raise their percentile ranks.

Conclusion

As an interviewer and a trainer, I can state that both the training and the implementation of the oral interview process has had a very positive effect on second language teachers. Though we conduct interviews not directly related to our local curriculum with students other than our own, we are afforded an experience that is not available within our own classrooms.

These tests are competency oriented and the vast majority of the students enrolled, limited or not in their speaking ability, realize that in order to be evaluated they need to talk. So talk they do.

Personally, I find interviewing the highest-level student the most difficult, as one has to extensively draw out vocabulary, structure, grammar, and other aspects in order to accurately judge the level. But at the same time, these students are the most interesting to talk to as they are usually most well-read on a variety of topics and more ready to communicate.

As far as nervousness is concerned, very few students have that problem. The students who are nervous are usually the very, very slow students and they will generally tell you when they enter the interview room, "I can't speak English but I understand everything." I do not think these students would do any better with a teacher they knew.

I certainly think it is a great opportunity for our students to be able to find out how competent they are in a second language. For some students, these interviews might be their answer to a career they are dreaming about. For others, just to know they can communicate in a second language will make them more emotionally secure in a new job or in a new English community. Therefore, wider horizons are opened to our students.

NEW BRUNSWICK SECOND LANGUAGE TESTING PROGRAM

GENERAL DESCRIPTION

The Second Language Testing Program is a service [illegible] the Department of Education for use [illegible]. The program will be avai[illegible] as part [illegible] second language programs in grades 10, 11 and 12 effective Spring 1976. It will replace the second language tests in the New Brunswick Senior High School Achievement Testing Program. The Second Language Testing Program will be administered to all second language students in participating schools during the spring of each year.

To make it possible for various instructional programs to be evaluated by the same testing program, the Second Language Testing Program tests will not be based on any particular manual or course of instruction. The tests will focus on proficiency in the second language.

The tests will cover the four basic skill areas of English and French as second languages. Reading and writing will be tested through paper-and-pencil tests, to be administered to students in grades 10, 11 and 12. Listening and speaking will be examined by means of language proficiency interviews, to be administered to students in grades 11 and 12.

It is planned that language interviews be recorded and the tapes retained at the Department of Education for a period of three years. Seniors would be given the opportunity to request that the recording of their own interview be sent to a prospective employer or to a post-secondary school which is considering their admission.

Scores obtained on the written tests and proficiency levels determined during the interviews will be reported officially during the summer of [illegible] along with scores on the Senior High Achievement Testing Program.

PAPER-AND-PENCIL TESTS

Multiple-choice tests will be used to test reading and writing. Fifty minutes will be allowed to complete each test.

THE READING TEST

The reading test will contain two types of questions (a) vocabulary in context and (b) reading comprehension, based on a variety of passages selected by New Brunswick teachers. Passages were selected and questions devised to cover a wide array of difficulties and content areas. The reading test will contain 60 questions.

THE WRITING TEST

The writing test will be an indirect measure of the writing skill. It will test the ability to distinguish among structures usually considered important in writing the second language and to select those appropriate for a given context. There will be three types of questions: (a) usage, (b) sentence correction, and (c) sentence completion. These questions will cover a variety of grammatical and stylistic problems and vary in difficulty. The writing test will contain 80 questions.

LANGUAGE PROFICIENCY INTERVIEWS

Language proficiency interviews will be conducted under standardized conditions by New Brunswick second language teachers who have been trained as second language interviewers through an in-service educational program implemented by the New Brunswick Department of Education and Educational Testing Service. Only those students in second language courses at grades 11 and 12 will be interviewed. Each language proficiency interview will give the student an opportunity to demonstrate, in a realistic conversational situation, the extent of his spoken mastery of the second language, as well as his ability to understand the spoken language. The specific content of the interviews will not be predetermined. *It will therefore not be useful for students to "prepare" for the interviews,* beyond engaging in similar conversational types of experiences. The following areas of proficiency will be evaluated: pronunciation, grammatical accuracy, vocabulary, fluency, and listening comprehension. A scale comprised of competency levels within each area of language proficiency will be employed. These scores will be tabulated for each student and summed according to a predetermined weighting. The sum will then be converted to a five-level overall language proficiency score.

OVERALL LEVELS OF LANGUAGE PROFICIENCY

Level 1: Able to satisfy travel needs and minimum courtesy requirements.

Level 2: Able to meet basic social demands and to satisfy simple needs related to school and work.

Level 3: Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical and social topics.

Level 4: Able to use the language fluently and accurately on all levels normally pertinent to the needs of all formal and informal conversations on practical, social, and work-related topics.

Level 5: Speaking proficiency equivalent to that of an educated native speaker.

NOUVEAU-BRUNSWICK PROGRAMME DES TESTS DE LANGUES SECONDES

DESCRIPTION GENERALE

Le Programme des Tests de Langues Secondes est un service du Ministère de l'Education mis à la disposition des écoles secondaires de la Province. Il sera mis en application dans ces écoles dès 1976, sur une base volontaire dans le cadre des programmes de langue seconde des 10e, 11e et 12e années. Ce nouveau programme remplacera les tests de langues secondes du Programme des Tests de Rendement Scolaire Secondaire, 2e Cycle. Le Programme des Tests de Langues Secondes sera administré au cours du printemps de chaque année à tous les élèves de langues secondes des écoles participantes.

Afin de permettre aux divers programmes d'enseignement d'être évalués par le même programme, le contenu du Programme des Tests de Langues Secondes ne sera pas basé sur un manuel de cours particulier ou sur un cours donné. Les tests mesureront la compétence dans la langue seconde.

[illegible] habiletés de base [illegible] langues secondes. [illegible] lecture et de [illegible] élèves de [illegible] 11e et 12e [illegible] l'expression [illegible] d'entrevues de compétence linguistique données aux élèves des 11e et 12e années.

Il est prévu que toutes ces entrevues de langues seront enregistrées et conservées pendant une période de trois ans au Ministère de l'Education. Les élèves de la 12e année auraient la faculté de demander que l'enregistrement de leur entrevue soit envoyé à un employeur potentiel ou à l'école post-secondaire où ils espèrent être admis.

Les notes obtenues aux tests écrits et les niveaux de compétence déterminés lors des entrevues seront communiqués officiellement pendant l'été de chaque année en même temps que les notes obtenues dans le cadre du Programme des Tests de Rendement Scolaire Secondaire, 2e Cycle.

LES TESTS ECRITS

Pour la lecture et la rédaction, on utilisera des tests à choix multiples. Les élèves disposeront de cinquante minutes pour finir chaque test.

TEST DE LECTURE

Le test de lecture comprendra des questions de deux genres: (a) vocabulaire en contexte et (b) compréhension de textes choisis par des enseignants du Nouveau-Brunswick. Les textes et les questions s'y rapportant représenteront plusieurs niveaux de difficulté. Le test de lecture contiendra 60 questions.

TEST DE REDACTION

Le test de rédaction sera une mesure indirecte de la capacité des élèves dans la rédaction. Il examinera leur habileté à distinguer parmi les structures généralement considérées importantes dans la rédaction de la langue seconde et à les choisir selon des contextes donnés. Il y aura trois genres de questions: (a) usage, (b) correction de phrases et (c) phrases à compléter. Ces questions auront plusieurs niveaux de difficulté et porteront sur une variété de problèmes grammaticaux et stylistiques. Le test de rédaction contiendra 80 questions.

LES ENTREVUES DE COMPETENCE LINQUISTIQUE

Les entrevues de compétence linguistique seront données, dans des conditions standardisées, par des enseignants du Nouveau-Brunswick qui ont été formés comme examinateurs de langues secondes au cours d'un programme de formation établi par le Ministère de l'Education et Educational Testing Service. Seuls les élèves des cours de langues secondes des 11e et 12e années auront des entrevues de langue dans le cadre du Programme des Tests de Langues Secondes. Chaque entrevue de compétence dans la langue donnera à l'élève l'occasion de démontrer, dans une conversation réelle et naturelle, le degré de son habileté à parler la langue seconde et à en comprendre l'expression orale. Le contenu précis des entrevues ne sera pas prédéterminé. *Il sera donc inutile que les élèves "se préparent" pour les entrevues*, si ce n'est de participer à des conversations semblables. Les domaines de compétence suivants seront examinés: prononciation, précision grammaticale, vocabulaire, facilité d'expression et compréhension auditive. Une échelle de compétence sera utilisée dans chaque domaine de compétence linguistique. Les notes seront établies pour chaque élève et additionnées selon un barème de coefficients prédéterminé. Le total sera ensuite converti en une note générale correspondant à un des cinq niveaux de compétence énumérés ci-après.

NIVEAUX GENERAUX DE COMPETENCE

Niveau 1: Peut satisfaire aux besoins simples du voyage et aux exigences minima de la courtoisie.

Niveau 2: Peut satisfaire aux exigences sociales de base et aux besoins simples se rapportant à l'école et au travail.

Niveau 3: Peut parler la langue avec suffisamment de précision structurale et lexique pour participer avec succès dans la plupart des conversations officielles ou ordinaires sur des sujets pratiques ou sociaux.

Niveau 4: Peut utiliser la langue couramment et avec précision à tous les niveaux des conversations courantes et spécialisées dans le domaine pratique ou social, ou touchant au travail.

Niveau 5: S'exprime avec une facilité égale à celle d'une personne instruite, native de la langue.

USING THE FSI INTERVIEW AS A DIAGNOSTIC EVALUATION INSTRUMENT

Stephen L. Graham

Brigham Young University

USING THE FSI INTERVIEW AS A DIAGNOSTIC EVALUATION INSTRUMENT

Stephen L. Graham

The Language Training Mission

The Language Training Mission (LTM) is located in Provo, Utah, adjacent to the Brigham Young University (BYU) campus. It was established to provide intensive language and cultural training for missionaries of the Church of Jesus Christ of Latter-day Saints (Mormon) who serve voluntary, two-year missions in many countries of the world.

Instruction began at the LTM in 1961 in Spanish and since that time has expanded to include Afrikaans, Cantonese, Danish, Dutch, Finnish, Flemish, French, German, Icelandic, Indonesian, Italian, Japanese, Korean, Mandarin, Navajo, Norwegian, Persian, Portuguese, Samoan, Serbo-Croatian, Swedish, Tahitian, Thai, and several Indian languages spoken in Latin America: Aymara, Cakchiquel, Guarani, Quechua, Quiche, and Quichua.

Five to six thousand missionaries are trained annually at the LTM in the languages mentioned above. The instructional staff is composed almost entirely of students at the university who are working their way through college. They are either native speakers of the languages or returned missionaries who have recently completed their missions and are at BYU pursuing their education. The number of language instructors at certain times during the year reaches as high as 300. Also included on the staff are 75 to 80 certified testers who conduct FSI interviews on a regular basis.

With the exception of approximately 100 missionaries a year who receive additional training, the missionaries learn one language and receive cultural training in an eight-week period of time. The missionaries are housed at the LTM and are required to speak their language for most activities during the day. This provides an ideal situation for total immersion in the language.

FSI Interview Adopted as Evaluation Instrument at LTM

Early in the spring of 1975 the Foreign Service Institute (FSI) interview was adopted as a major evaluation instrument at the LTM to help determine the overall language proficiency of the missionaries going through the program. There were three main reasons for its adoption: (1) the FSI interview is a well-designed, well-respected instrument, and provides a means of comparing results in oral language proficiency with other language institutions; (2) it is relatively simple to administer across different languages and, with periodic in-service workshops, quality control can be maintained within and across languages; (3) the "interview setting" is ideal for giving immediate, individually tailored feedback to the person being interviewed.

Protase E. Woodford of Educational Testing Service (ETS) conducted the initial training for the first team of Spanish testers early in 1975, and by mid-January 1977 certified testers had been trained in the twenty-one languages being taught at that time. Regular seminars for retraining and in-service workshops have since continued, including a two-day seminar in August 1977 that was given by John L. D. Clark of ETS.

Upon arrival at the LTM, missionaries who have had prior experience in their target language receive an "entering FSI interview." All missionaries, without exception, receive a "departing FSI interview" at the conclusion of their LTM stay. Those who desire interim interviews for diagnostic purposes have this option available to them at any time during their stay. Scores are not recorded for the interim interviews; the emphasis is on giving useful feedback.

FSI Tester Training at the LTM

The training of FSI testers at the LTM is conducted in three segments: (1) acquiring rating skills, (2) acquiring interviewing skills, and (3) in-service and retraining to maintain those skills.

Rater training is provided through a self-instructional package entitled "Oral Language Proficiency Test Training Manual" (Part A), prepared by the LTM. The manual is accompanied by several sets of practice tapes (prerecorded and prerated FSI interviews) and a set of certification tapes. The trainee checks out the materials and works through them at a comfortable rate for him. The practice tapes give him an opportunity to practice his rating skills by assigning ratings to actual prerecorded interviews and then comparing his ratings with those of experienced testers.

To move ahead into the training program for interviewing skills, the trainee must correctly assign FSI ratings for the prerecorded interviews of the certification tapes.

Interview training is provided on an individual basis as well. Each trainee works in an apprentice-type situation where he receives personal, on-the-job training from an experienced tester. He begins by watching interviews that have been videotaped and by observing live interviews conducted by the experienced tester. The trainee then begins participating in actual interviews until he feels confident in conducting an effective interview on his own.

The emphasis of the interview training is to ensure that the tester provides a comfortable atmosphere in which the missionary is able to perform at his maximum capacity in the language.

In-service workshops are conducted every two months to provide follow-up training and remedial help where needed in both rating and interviewing skills. Activities of the workshops consist of conducting actual interviews on the spot and rating interviews that have been prerecorded on audio and video cassettes. Ratings are assigned independently by each tester and the results are then discussed as a group. Testing teams representing all languages taught at the LTM are present at the workshops.

English is used for all initial training and workshop sessions. This does have some disadvantages in that the majority of testers are not native English speakers, but it helps maintain quality control across languages. Using English also helps keep the focus of the workshops on rating a person's ability to perform certain tasks in the language and avoids the myriad "linguistic" concerns that sometimes are raised when dealing with so many different languages.

In-House Evaluation of FSI Testing Program at LTM

At the close of 1977 (the first full year of FSI testing in all languages taught at the LTM), the administrative staff conducted an informal, in-house evaluation of the FSI testing program. This was to determine how well the program was fulfilling the three main purposes for which it was adopted. At the conclusion of the evaluation, the staff was encouraged by the quality and consistency of the testing results. Concern was expressed, however, about its usefulness in providing helpful feedback to the missionaries. A summary of the evaluation results follows:

During 1977, 6,193 FSI interviews were conducted in twenty-four languages. This number includes both "entering" and "departing" interviews. Of a randomly selected 763 interviews conducted in French, German, Japanese, and Spanish between the months of January and June 1977, there were only 156 discrepancies between independent ratings assigned by the interviewer and the rater before consultation. Of those 156 discrepancies, 155 were no larger than a "plus." In other words, LTM FSI testers in these four languages agreed on the exact ratings 92.7 percent of the time without consulting each other. In the few cases where there were disagreements, the difference was rarely more than a "plus."

The reliability of ratings across languages is a topic of every bimonthly workshop. As mentioned earlier, the majority of testers are not native English speakers. All training on this level, however, is conducted in English expressly for the purpose of ensuring consistency across languages. This is done by having all testers independently rate prerecorded interviews from a variety of sources. For example, ETS recordings are frequently used, along with those prerecorded by various teams represented at the workshops.

As an example of tester performance during these regular workshops, the results of the most recent one, held in February of this year, are of interest: Of 102 independent ratings assigned during the workshop prior to consultation, 95 were in agreement, with only 7 ratings being either a "plus" too high or too low. Several of the interviews used for rating during the workshop were prerecorded on audio cassettes, others were recorded on video cassettes, and one interview was conducted live.

The results of the evaluation up to this point indicated to the administrative staff that the general operation of the FSI testing program was improving both within and across languages. They also showed that the initial and in-service training programs for testers had become systematic and quite effective.

In addition to having a smoothly functioning FSI testing program with adequate training for personnel and reliable ratings, one of the goals of the administrative staff is to provide missionaries with as much diagnostic help as possible during their LTM stay. This should enable them to increase their language proficiency significantly before leaving for the countries to which they are assigned.

During February 1978, feedback was elicited from language instructors, testers, and missionaries to determine the general feeling about how much diagnostic help was actually being given. Three weaknesses were consistently mentioned and confirmed by observing actual interviews. These weaknesses were:

1. Lack of sufficient time to follow up on deficiencies. (Most of the interviews are given to missionaries three or four days prior to their departure for the assigned countries.)

2. Lack of a systematic procedure for the tester to organize the feedback in a usable format for the missionary.

3. Lack of a systematic procedure for getting the feedback back into the instructional program and ensuring that problems are remedied as well as diagnosed.

Procedures for Providing Systematic Diagnostic Feedback

In an effort to facilitate the flow of useful, systematic feedback both to the individual missionary and into the instructional program itself, the following changes and modifications are proposed:

1. The FSI "entering" and "departing" interviews will no longer be conducted for every missionary. They will be conducted, rather, on a random selection basis to provide the administrative staff with a continual flow of statistical data for purposes of evaluation.

2. Each missionary will receive an interim diagnostic FSI interview during the third and sixth weeks of his stay at the LTM. These interviews will be conducted in the same manner as the regular FSI interview, except that diagnostic feedback will be given to the missionaries in lieu of FSI ratings.

3. Testers will be provided with a diagnostic feedback checklist specific to their language. This sheet will be used to record patterns of deficiencies in a missionary's speech during the interview. The form will be prepared in triplicate. At the conclusion of the interview one copy will be given to the missionary for his own personal reference, one copy will be sent to the instructional staff, and one will be retained in the testing center.

This form will provide a means for the instructional staff to watch for high-frequency items indicating specific areas of deficiency unique to that particular language. Mini-classes will then be conducted during the personal study time of the missionaires, and the most common errors in grammar principles, vocabulary, comprehension, fluency, and pronunciation will be treated on an individual and a group basis. (An example of the French diagnostic feedback sheet is included as Appendix A.)

Conclusions

The administrative staff feels these modifications in procedures will greatly enhance the usefulness of the FSI interview in a practical way without changing the test itself or the purposes for which it was designed. It is important to the Language Training Mission to be able to compare results in oral language proficiency with other language institutions.

It is expected that the diagnostic feedback sheet will need periodic revision and modification with respect to both scope and layout. These changes will be made as needed over the next few months in a trial run. The idea, however, of taking full advantage of the "interview setting" for giving personal, oral feedback to individuals is the intent of the recommended changes. The emphasis on "oral evaluation" is especially important at the LTM, where the emphasis in language training is on acquiring speaking and listening comprehension skills.

The FSI interview testing program (both diagnostic and traditional), accompanied by the traditional written testing program, will provide the LTM with useful formative and summative evaluation data. Both are essential to ensure individual improvement for the missionaries and to upgrade and modify instructional programs and materials.

Appendix A

FSI DIAGNOSTIC FEEDBACK - FRENCH

NAME ______________________ DATE ______________________ INTERVIEWER ______________________

CHECK (✓) THE ITEMS BELOW WITH WHICH THE PERSON BEING INTERVIEWED HAS DIFFICULTY. MAKE ADDITIONAL COMMENT AND OBSERVATIONS IN THE SPACE PROVIDED. THIS EVALUATION WILL BE USED IN GIVING REMEDIAL HELP TO THE PERSO BEING INTERVIEWED TO INCREASE HIS PROFICIENCY.

GRAMMAR

___speaks in infinitive or with no verbs at all

___preposition *à*

___preposition *de*

___*à + le, les = au, aux*

___*de + le, les = du, des*

___*avoir* and *être* as auxiliaries

___direct object pronouns *le, la, les*

___indirect object pronouns *lui, leur*

___*Il y a* and *depuis* used with time

___*en* and *dans* used with time

___*c'est* and *il est*

___definite article as in *la charité*

___*ne...jamais, ne...personne*, etc.

___prepositions of place *à*, *en* and *au*

___*de* as in "*je n'ai pas de...*"

___adverbs vs. adjectives as with *correct* and *correctement*

___reverts to English word order

___verb endings

___present tense

___past compound tense

___future tense

___imperfect tense

___conditional tense

COMMENTS:

___subjunctive mood

___adjective agreement

___agreement of past participle

VOCABULARY

___*connaître* vs. *savoir*

___*parler* vs. *dire*

___*bien* vs. *bon*

___*peuple* vs. *gens*

___*fois, temps, l'heure, moment*

___*d'accord* vs. "OK"

___*pour* and *pendant* and time

___*changer, changer de*, and *changement*

___*avant de* vs. *avant que* ...

___*après avoir* vs. *après être* ...

___*vous* and *tu* vs. impersonal *on*

___*plus, très*, and *trop*

___*mieux* vs. *meilleur*

Makes the following common errors:

___*attendre "pour"*

___*chercher "pour"*

___*"pour, sur"*

___*bènir "avec"*

___*avoir besoin "pour"*

___*"plus" mieux*

COMMENTS:

___Gropes for specific, appropriate words or expressions when discussing his area of expertise (missionary related topics).

___Cannot describe objects, feelings, and situations when he does not know the specific word or expression.

LTM 6/26/78

FSI DIAGNOSTIC FEEDBACK - FRENCH (CONT.)

COMPREHENSION

___When the interviewer spoke at his normal speaking speed the interviewee had difficulty following him.

___When the interviewer spoke on general topics other than those very familiar to the interviewee, the latter understood isolated words and expressions but generally did not understand the full context of ideas.

1. When the interviewer spoke on the topic of ______________, the interviewee ________________ ____.

2. When the interviewer spoke on the topic of ______________, the interviewee ________________ ____.

3. When the interviewer spoke on the topic of ______________, the interviewee ________________ ____.

4. When the interviewer used the word (or expression)____________, the interviewee __________ ____.

5. When the interviewer used the word (or expression)____________, the interviewee __________ ____.

6. When the interviewer used the word (or expression)____________, the interviewee __________ ____.

FLUENCY

The items which are checked below describe the fluency of the interviewee's language:

___Has difficulty speaking at his own natural speaking speed	___Speaks at natural speed
___Pauses are unnatural and illogically placed	___Pauses natural and logical
___Speech is irritating and annoying to listen to over long period of time	___Speech not annoying
___Phrases are broken and incomplete	___Phrases smooth and complete
___Speaking generally requires a great effort on the part of the interviewee	___Speech is effortless

PRONUNCIATION

The person being interviewed has difficulty with the items below which are checked:

___i as in __________, __________, __________.

___e as in __________, __________, __________.

___ɛ as in __________, __________, __________.

___ə as in __________, __________, __________.

___y as in __________, __________, __________.

___ɔ (open) as in ______________, __________.

___o (closed) as in ____________, __________.

___u as in __________, __________, __________.

Liasons:

optional as in ______________, __________.

obligatory as in ____________, __________.

prohibited as in ____________, __________.

Nazals:

___an / am as in __________, __________, __________.

___in / im as in __________, __________, __________.

___on / om as in __________, __________, __________.

___un as in ______________, __________, __________.

___"ss" between vowels: *impression* ______________.

___"s" between vowel and consonant: *enthousiasme*.

___"e" as in *se lever, revenir, devez,* ____________.

___i / e before double consonant as in *innocent, ennemi*, ______________, __________, __________.

___o before double consonant as in *bonne, occuper, possible, pomme*, ________________, __________.

COMMENTS:

____

____

____

DIRECT TESTING OF SPEAKING SKILLS
IN A CRITERION-REFERENCED MODE

Robert B. Franco

Defense Language Institute

DIRECT TESTING OF SPEAKING SKILLS IN A CRITERION-REFERENCED MODEl

Robert B. Franco

Background

The Defense Language Institute (DLI) and its predecessor, the Army Language School (ALS), have traditionally emphasized the development of oral skills in their foreign language programs. Although in the past few years other primary objectives, of a military-technical nature, have been pursued, the main emphasis has remained on developing speakers of foreign languages to an S-3 level of proficiency. Ironically, the speaking skills have been the most elusive and difficult to measure with a satisfactory degree of objectivity.

Historical Perspective

At DLI the search for an effective system of evaluation of speaking skills can be traced back to the days of the Army Language School and extends until the present time, but for the purposes of this paper, the period will be divided into pre-1976 and post-1976 segments. In our pre-1976 couses, the core of the lesson unit was a "basic dialog," charged with presenting certain grammatical features within the context of a high-frequency, authentic situation. Traditionally, the dialog was introduced in class, then studied until "fully understood" and memorized at home. The next day, the dialog was reviewed and enacted in the classroom, as realistically as possible. A good imitation by the student of the native model's pronunciation and fluency, an indication of a clear understanding of what was being said, plus the native-like use of important paralinguistic features, constituted the evaluation criteria.

The acceptability of the student's performance depended on the powers of observation and the subjective appreciation of the instructor. Furthermore, an acceptable performance in class was recognized as sufficient proof of the student's capacity to perform effectively on the job.

Cognizant of the subjectivity that permeated this method of evaluating speaking skills, ALS/DLI instituted a less informal system, which included weekly, monthly, and final oral examinations. The weekly tests consisted of a series of questions based on the materials covered during that week. These questions were read aloud by the instructor, who then noted the accuracy and completeness of the student's responses. For the monthly and final examinations, one or two bilingual conversations were

[1] The views of the author do not purport to reflect the position of the Department of the Army or the Department of Defense.

added in which the examinee played the impromptu role of interpreter. Notes and tallies were kept, but the scoring was still based on a subjective appreciation of the examinee's performance, even when an examiner other than the classroom teacher was the scorer. As part of the system, the oral score was computed with the scores of pencil-and-paper tests given for other skills, and a composite of all test scores was then computed with the average of the daily grades for the testing period.

Somehow, our good teaching survived our poor testing, at least within our system. To illustrate, in 1973 we took a ten-year block of these composite scores of approximately 1,000 Spanish basic course students and compared the scores with those obtained by the same students on the listening comprehension part of the Defense Language Proficiency Test. To our surprise, a correlation of .91 was discovered, although the correlation for other languages is about .60. This relieved us momentarily, but of course did not validate our system.

In the late fifties and early sixties, our expectations were raised by the development and refinement of the Foreign Service Institute (FSI) "techniques for the testing of speaking proficiency," followed by publication of the Modern Language Assocation (MLA) Cooperative Foreign Language Tests and the Modern Language Association Proficiency Tests for Teachers and Advanced Students. DLI examined the new instruments very carefully, tried them out, and adopted their formats with the modifications required by the nature of our student population and their special needs.

For the pre-1976 Spanish basic course, specifically, we adopted the FSI model and used it, experimentally, as a proficiency, placement, and achievement test. However, its full utilization was inhibited by two factors: the limited scope of our basic course (with a final objective of S-3) and the absence in the course design of interim objectives that would have addressed the S-1 and S-2 levels chronologically and permitted diagnostic use of the structured oral interview based on FSI techniques.

We found the MLA speaking tests were not as readily adaptable to the Spanish basic course, mainly because the tests had a different content and employed techniques with which our examinees were not as familiar. As with the FSI interview, the internal structure of the course was also an inhibiting factor, although this was later remedied in the new course design. Features of the MLA model, nevertheless, were incorporated into the "level tests" developed by DLI and Educational Testing Service.

The New Spanish Basic Course, Post-1976

In the mid-seventies, a new DLI Spanish basic course was designed and developed under the growing influence of a criterion-referenced instruction (CRI) approach, derived from the Interservice Procedures for Instructional Systems Development (IPISD), a model produced by the Florida State University under a joint interservice contract. Thus, a system designed primarily for military instruction was transplanted into the foreign language curriculum.

In addition to this CRI general orientation, the new design addressed the sequential achievement of skill levels I and II as interim objectives, keeping skill level III as the final objective of the basic course. Schematic diagrams for the pre- and post-1976 course design are shown in Appendix A.

Course Design

The course consists of nine general modules and one enrichment/ remedial module, to be covered in no longer than twenty-seven weeks. Modules 1, 2, and 3 address skill level I; modules 1 through 6, with emphasis on 4, 5, and 6, address skill level II; and all nine modules, with emphasis on 7, 8, and 9, aim at skill level III. The evaluation track includes nine module tests and three level tests, with the level 3 test complemented by a comprehensive achievement test, the Defense Language Proficiency Test (DLPT), and a structured oral interview, limited to skill level III. In addition, each of the six lesson units in a module contains a series of criterion checks for the evaluation of stated lesson objectives, with emphasis on the communication frame to check speaking ability. A separate track of criterion-referenced checks evaluates listening comprehension skills.

New Evaluation Design

The field test of the materials indicated the need to consolidate the various types of tests into a comprehensive, criterion-referenced evaluation track.

The new track combined the best features selected from each of the previous components. This selection was based primarily on student and faculty input that was, admittedly, personal and subjective. The result was a battery of partly norm-referenced and partly criterion-referenced tests called Comprehensive Hybrid Achievement Tests (CHATs). Our new technology, however, required a clearer CRI orientation, so we reexamined the objectives and the criteria, and adjusted the instruments. This produced the present Major Criterion-Referenced Tests (MCRTs): Anchor CRT 1, Anchor CRT 2, and Final CRT, which evaluate the attainment of the objectives assigned to skill levels I, II, and III, respectively. Neither the module tests, the lesson unit quizzes, nor the listening comprehension CRTs were modified, but closer coordination was recommended of lesson objectives, communication frames, and the speaking MCRTs.

The MCRTs test seven component skills independently. Speaking is listed, arbitrarily, as number IV. The content outline for the complete MCRT battery is shown in Appendix B.

The Speaking MCRTs

Specifications. A complete set of specifications for the Spanish Speaking MCRTs is included in Appendix C.

Format. The speaking test consists of a two-part oral interview between an examinee and one specially trained native speaker in Spanish. The first part of the interview is related to specific topical areas about which the examinee has knowledge. Spoken Spanish responses by the examinee are elicited by spoken Spanish questions or statements by the interviewer and systematically based upon the list of topics.

The second part of the test is conducted in the same manner. Instead of topics, role-playing situations are utilized to form the basis for the examinee's responses. Both topical areas and role-playing scenarios are printed in English in the test booklet that is given to the examinee at the beginning of study for the modules to be tested. A separate booklet is provided for the interviewer to provide the information necessary to prepare, conduct, and score the interview.

During the study of the modules to be tested, the student is encouraged to act out the scenarios pertinent to each lesson and to be checked out by his or her instructors. In fact, the students themselves have developed a check sheet for each role-playing situation and concentrate their attention on those scenarios that are not specifically covered in the communication frames of the lesson CRT.

Content. As stated earlier, the Spanish MCRTs parallel the objectives and content of the basic course. Anchor CRT 1, for example, addresses tasks derived from the definition of skill level I in speaking that correspond to the speaking objectives of modules 1, 2 and 3, which are the targets of the test.

To illustrate:

Level I objectives (S-1 tasks):

1. Use greetings and leave-taking expressions. Offer apologies.

2. Make simple social introductions of self and others.

3. Ask and tell time of day, day of week, date.

4. Order a "simple" meal.

And so forth.

Elements of task 4, for instance, have been assigned to lesson 7 as its speaking objective, within the format and criteria of effective role-playing of restaurant scenarios. To verify the achievement of this objective, after all enabling objectives have been satisfied, the student is tested in the four role-playing situations of the communication frame, which is the lesson's speaking test. Also, while working in the first three modules of the course, the student prepares and is checked out on the six role-playing situations included in Anchor CRT 1 for task 4. Thus, when the test is formally administered, a passing score on any of the six scenarios would satisfy the requirements of this task.

This close parallelism may constitute one of the best features of the Spanish MCRTs.

Administration. The test is administered in the form of a structured oral interview. The interviewer must be a native speaker of Spanish and specially trained to use this technique. Though structured, each interview is unique. For this reason, standardized alternate test forms employed for measuring the other skills in the Spanish MCRT series are not used in the speaking test.

Separate guides have been prepared for Anchor CRT 1, Anchor CRT 2, and the final examination, each with examiner's and examinee's versions.

a. Examiner's guide. Each guide provides detailed information on the procedures to be followed and supplies the topical and situational information that give the examination its structured elements. It is essential that interviewers administering these speaking tests be thoroughly familiar with the contents of both the examiner's guide and the examinee's guide.

b. Examinee's guide. Each examiner's guide has a companion examinee's guide. The guide for the examinee provides procedural, topical, and situational information and is given to the student when he or she begins study of the modules with which each guide is associated. The student is instructed to become familiar with the contents of the guide and to bring the guide to the test site. Each guide also contains a removable student rating sheet. Its use will be described in the section about scoring.

c. Time allocation. Time allowed for administration of the speaking tests is indicated in the examinee's guide, the examiner's guide, and in table 1 of the administration and scoring manual prepared for the MCRTs, as shown in Appendix D.

d. Observers. The Spanish MCRTs are designed for use in a face-to-face, one examinee/one interviewer situation. The presence of an independent scorer, an observer, or an interviewer trainee is permitted. Any such third person present during the interview must remain silent and unobtrusive.

e. Recording. Recording the oral interview is permitted. These recordings may be used for independent scoring, training interviewers, or rating interviewee performance by another rater. Most reel-to-reel and cassette tape recorders have only a single microphone input jack. For this reason, the microphone must be carefully placed so both interviewer and examinee voices will be recorded. Preadjusting the equipment under actual test conditions is recommended.

Scoring. The speaking tests may only be scored by trained scorers who have expert knowledge of the Spanish language. Full details on scoring the speaking tests are contained in the examiner's guides that have been prepared for each Anchor speaking test and the final speaking examination. Since no two interviews are conducted identically and examinee responses can vary, the speaking test is not arranged in standardized alternate forms. A separate rating scale has been prepared for each Anchor test and for the final test. Appendix E shows the student rating sheet for Anchor CRT 2. Similar sheets (with different rating level weights and percentage conversion tables) have also been prepared for Anchor CRT 1 and the final examination. While speaking is subject to minimum acceptable performance standards, a special provision has been added to these tests so that examinee performance can also be expressed as a performance skill level.

a. Ratings. Performance ratings are used to derive skill points from which the score is determined. The procedure is the same for the Anchor tests and the final examination. A three-point rating scale is applied to five linguistic categories in accordance with the statements of performance criteria. The ratings based upon the examinee's performance are not language skill levels, but points from which to derive a score. It is this point score that can be converted to conventional language skill levels, to percentage grades, or to pass/fail grades. A separate rating sheet is provided for each speaking test to reflect slightly different weights for certain linguistic categories. The procedure for using the rating sheets is the same for all speaking tests.

b. Computation of Points and Score Conversions. The examiner is required to use the following procedure:

1. Using the computation table at the top of the rating sheet, judge the examinee's performance on each of the five linguistic categories, determine the number of points derived by using the appropriate rating column (1, 2, or 3), and enter that number of points in the space provided under "Skill Points." Add the column of skill points. This produces the examinee's point score.

2. The score-to-level conversion table is located at the lower left-hand side of the rating sheet. Using the total number of points scored, circle the appropriate level opposite that band of scores. Enter the skill level attained in the space marked "Skill Level" at the bottom of the page.

3. The score-to-percentage conversion table is located at the lower right-hand side of the rating sheet. Using the total number of points scored, circle the appropriate percentage score for points scored. Enter the percentage score attained in the space marked "Percentage Score" at the bottom of the page.

4. Based upon the minimum acceptable performance standard for speaking, check the "Pass" or "Fail" block at the bottom of the page. The criteria for each linguistic category were adapted from the definitions previously used at DLI, derived primarily from the FSI interview materials. Performance criteria for Anchor CRT 2 are reproduced in Appendix F.

Validation

The components of the Spanish MCRT battery were produced between July 1976 and November 1977, and, on the assurance of subject matter experts, these MCRTs are considered validated by DLI and are being monitored to ensure that they continue to meet design criteria (the concept of "internal validation" vs. "external validation"). By February 10, 1978, the tests had been administered to only 111 students, with the following basic results:

MCRT 1	N = 50	Passed = 45	Failed = 5
MCRT 2	N = 46	Passed = 43	Failed = 3
	(N = 15*)	(13)	(2)
MCRT3	N = 15*	Passed = 14	Failed = 1

*These students (Class 01LA24W 0977) were not administered CRT 1, because the test was not available when the class reached the S-1 level.

Admittedly, this is too small a sample to ensure utility for external uses, but it is considered sufficient for DLI purposes. Furthermore, the initial reaction from both examinees and examiners is encouraging. Following are a few of the comments gathered to date about the test:

> "It measures the functional competences stated as learning objectives."
>
> "Both the limited scope of each CRT and its use of content-sensitive scenarios tend to guarantee a fuller exploration of the stated objectives [than is true of other tests used previously]."
>
> "The student is encouraged to be checked out by the instructor on each of the interview topics and role-playing scenarios one by one, and to use this informal appraisal of his or her performance diagnostically for immediate remediation."

"Role-playing is preplanned, integrated into the course, and is not a surprise at the time of the test."

"Because of scope limitations, no exploratory time is required, greatly reducing administration time, especially for MCRTs 1 and 2."

"Examiners must use the student rating sheets to assign S-ratings and other scores. Thus, 'experienced judgment' plays a lesser role, which tends to reduce the subjectivity of the scoring system."

"The tests appear to have 'inherited' the validity of the FSI interview, and could perhaps surpass it."

These opinions will be corroborated or disclaimed through our mediation and monitoring procedures. Meanwhile, several test features have been identified for critical evaluation, for example:

The 70 percent minimal acceptable performance cutoff. (This was set by the user agencies, but the test developers feel it could be raised, to better equate test performance with on-the-job performance requirements.)

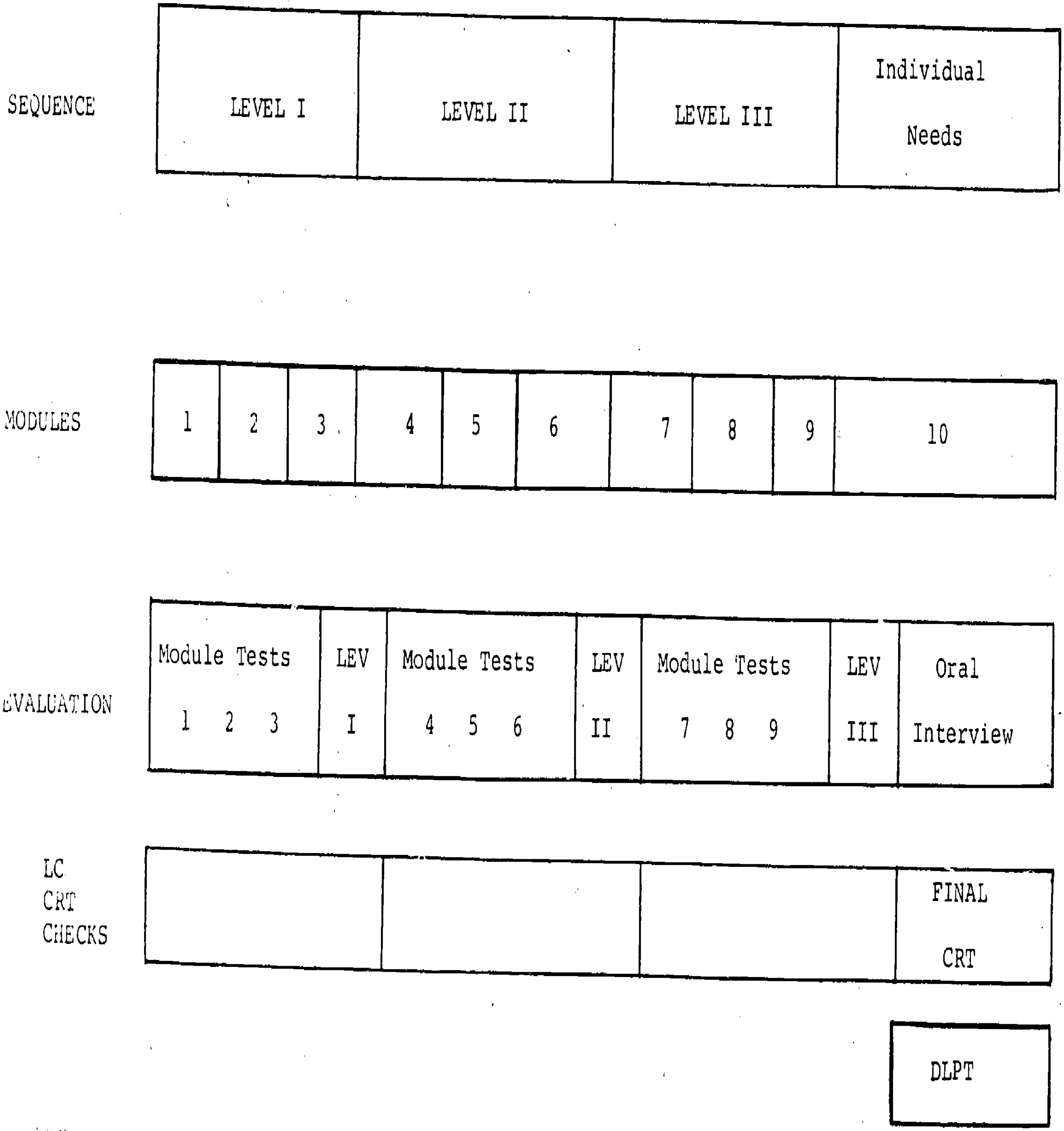
The number of role-playing scenarios and the procedures used for the selection of those actually tested. (The procedure could include the examiner's review of the examinee's record of scenarios checked out, and of any specific job requirements known.)

The "up-to-date" situational orientation of the interview and the role-playing scenarios. (Specific changes in course objectives dictated by changing conditions in the field will affect test content.)

Conclusions

It has been apparent to the developers of the Spanish MCRTs that both examiners and examinees approve of the speaking tests. We have observed in the students an attitude of enthusiasm and a sincere desire to prepare fully for the tests and to excel in their performance. There seems to be no doubt as to the content validity of the tests. As for their predictive validity, the criterion-referenced ambiance in which the tests are used and our informal observation of the initial results provide us with encouragement. Nevertheless, in the absence of sufficient data, no final conclusions can be made at this time on the overall efficacy of the DLI Spanish speaking tests as criterion-referenced instruments. As we gather data and develop supportive conclusions, we shall be happy to share them with any interested persons.

SPANISH BASIC COURSE DESIGN - 1975



SEQUENCE: LEVEL I | LEVEL II | LEVEL III | Individual Needs

MODULES: 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10

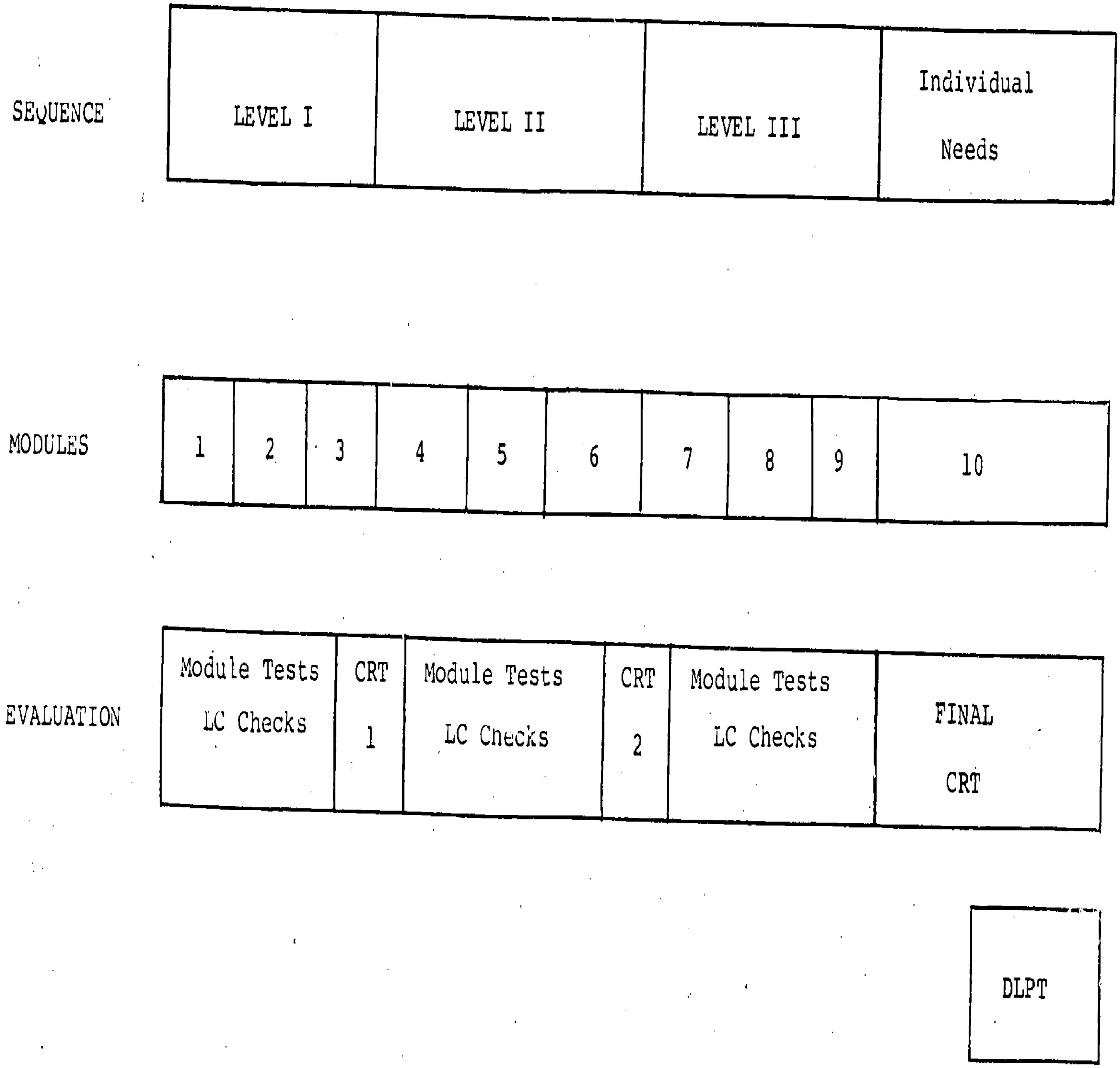
EVALUATION: Module Tests 1 2 3 | LEV I | Module Tests 4 5 6 | LEV II | Module Tests 7 8 9 | LEV III | Oral Interview

LC CRT CHECKS: | | | FINAL CRT

DLPT

Appendix A

SPANISH BASIC COURSE DESIGN - 1976



SEQUENCE: LEVEL I | LEVEL II | LEVEL III | Individual Needs

MODULES: 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10

EVALUATION: Module Tests LC Checks | CRT 1 | Module Tests LC Checks | CRT 2 | Module Tests LC Checks | FINAL CRT

DLPT

Appendix B

SPANISH MAJOR CRTs - DESIGN

	CRT #1	#2	FINAL	CRT #1	#2	FINAL	C-R
I LISTENING COMPREHENSION							
Conversations	(3,	3,	3)	10	10	25 M/C Items	
Broadcasts	(3,	3,	3)	10	10	25 M/C Items	
			Total =	20	20	50 M/C Items	70%
II READING COMPREHENSION							
Signs	(3,	3,	-)	3	3	- M/C Items	
Notices	(3,	3,	-)	7	7	- M/C Items	
Headlines	(3,	3,	-)	3	3	- M/C Items	
Articles	(2,	2,	6)	7	7	50 M/C Items	
			Total =	20	20	50 M/C Items	70%
				15	20	50 Minutes	
III TRANSLATION							
Text (100, 150, 200 words)				20	30	40 Key Words	70%
(Lexical Aids)				15	30	45 Minutes	
IV SPEAKING						S-1, S-2, S-3	
1- Interview/Conversation				5	5	10 Minutes	70%
2- Role Playing (2, 3, 4 Sits.)				10	15	20 Minutes	
V WRITING							
1- Completion				12	24	36 Items	70%
2- Transformation				6	12	18 Items	70%
3- Composition				1	2	3 Comps.	70%
				20	45	60 Minutes	
VI NUMBER TRANSCRIPTION							
1- Five 10-Number Series				(3	4	5 digits)	90%
2- Ten In-Context Numbers				(Card., Ord. & Fract.)			90%
VII GENERAL TRANSCRIPTION							
Conversations	(3,	3,	3)	60	90	135 Minutes	87.5%
Broadcasts	(3,	3,	3)				

Appendix C

TARGET LANGUAGE CRITERION-REFERENCED TEST ANCHOR CRT I SPECIFICATIONS December 1976

Speaking

The speaking test is divided into two parts: Part 1, in a direct conversation/interview format, and Part 2, in a role-playing format.

1. Part 1/Stimulus and Task - Given not less than 15 oral questions sequenced into an informal conversation, covering at least 3 separate Basic Topics from those listed in the Examiner's Guide, and presented orally by the interviewer, the examinee will answer the questions orally, as completely and fluently as possible.

2. Part 1/Conditions -

a. The conversation/interview will utilize not more than 5 minutes of the 15 minutes allocated to the speaking test.

b. No lexical aids are permitted.

c. Vocabulary and grammatical features used in the stimulus must be limited to those covered in the course of instruction for which the examinee is being measured.

3. Part 1/Criterion -

a. Scoring is accomplished by the interviewer by keeping mental notes or casually noting on the Student Rating Sheet the level of ability demonstrated by the examinee on each sub-skill.

b. After both Part 1 and Part 2 have been completed, the examiner combines his/her observations into one grade for each ability and computes the raw score using the S-1 COMPUTATION TABLE. (The computation table and scoring procedures are provided in the Examiner's Guide.)

c. No criterion is prescribed for Part 1, but a 30 raw-score cut-off (equivalent to an S-1 Level) is established for the entire speaking test.

4. Part 2/Stimulus and Task - Given not less than three role-playing scenarios selected as recommended in the Examiner's Guide, the examinee will assume the roles indicated in the scenarios and conduct them with the instructor as naturally and fluently as possible.

5. Part 2/Conditions -

a. The three scenarios must be completed within 10 minutes.

b. The examinee is permitted to quickly read the instructions for the scenario, but the use of lexical aids is not permitted.

c. Vocabulary and grammatical features used in the stimulus must be limited to those covered in the course of instruction for which the examinee is being measured.

6. Part 2/Criterion -

a. Scoring is done as described in 3a and b above.

b. No criterion is prescribed for Part 2, but a 30 raw-score cut-off (equivalent to an S-1 Level) is established for the entire speaking test.

TARGET LANGUAGE CRITERION-REFERENCED TEST
ANCHOR CRT II SPECIFICATIONS
December 1976

Speaking

The speaking test is divided into two parts: Part 1, in a direct conversation/interview format, and Part 2, in a role-playing format.

1. Part 1/Stimulus and Task - Given not less than 15 oral questions sequenced into an informal conversation, covering at least 3 separate Basic Topics from those listed in the Examiner's Guide, and presented orally by the interviewer, the examinee will answer the questions orally, as completely and fluently as possible.

2. Part 1/Conditions -

a. The conversation/interview will utilize not more than 5 minutes of the 20 minutes allocated to the speaking test.

b. No lexical aids are permitted.

c. Vocabulary and grammatical features used in the stimulus must be limited to those covered in the course of instruction for which the examinee is being measured.

3. Part 1/Criterion -

a. Scoring is accomplished by the interviewer by keeping mental notes or casually noting on the Student Rating Sheet the level of ability demonstrated by the examinee on each sub-skill.

b. After both Part 1 and Part 2 have been completed, the examiner comb nes his/her observations into one grade for each ability and computes the raw score using the S-2 COMPUTATION TABLE. (The computation table and scoring procedures are provided in the Examiner's Guide.)

c. No criterion is prescribed for Part 1, but a 45 raw-score cut-off (equivalent to an S-2 Level) is established for the entire speaking test.

4. Part 2/Stimulus and Task - Given not less than four role-playing scenarios selected as recommended in the Examiner's Guide, the examinee will assume the roles indicated in the scenarios and conduct them with the instructor as naturally and fluently as possible.

5. Part 2/Conditions -

a. The four scenarios must be completed within 15 minutes.

b. The examinee is permitted to quickly read the instructions for the scenario, but the use of lexical aids is not permitted.

c. Vocabulary and grammatical features used in the stimulus must be limited to those covered in the course of instruction for which the examinee is being measured.

6. Part 2/Criterion -

a. Scoring is done as described in 3a and b above.

b. No criterion is prescribed for Part 2, but a 45 raw-score cut-off (equivalent to an S-2 Level) is established for the entire speaking test.

TARGET LANGUAGE CRITERION-REFERENCED TEST FINAL EXAMINATION SPECIFICATIONS December 1976

Speaking

The speaking test is divided into two parts: Part 1, in a direct conversation/interview format, and Part 2, in a role-playing format.

1. Part 1/Stimulus and Task - Given not less than 15 oral questions sequenced into an informal conversation, covering at least 3 separate Basic Topics from those listed in the Examiner's Guide, and presented orally by the interviewer, the examinee will answer the questions orally, as completely and fluently as possible.

2. Part 1/Conditions -

a. The conversation/interview will utilize not more than 10 minutes of the 30 minutes allocated to the speaking test.

b. No lexical aids are permitted.

c. Vocabulary and grammatical features used in the stimulus must be limited to those covered in the course of inst-uction for which the examinee is being measured.

3. Part 1/Criterion -

a. Scoring is accomplished by the interviewer by keeping mental notes or casually ncting on the Student Rating Sheet the level of ability demonstrated by the examinee on each sub-skill.

b. After both Part 1 and Part 2 have been completed, the examiner combines his/her observations into one grade for each ability and computes the raw score using the S-3 COMPUTATION TABLE. (The computatation table and scoring procedures are provided in the Examiner's Guide.)

c. No criterion is prescribed for Part 1, but a 63 raw-score cut-off (equivalent to an S-3 Level) is established for the entire Speaking test.

4. Part 2/Stimulus and Task - Given not less than four role-playing scenarios selected as recommended in the Examiner's Guide, the examinee will assume the roles indicated in the scenarios and conduct them with the instructor as naturally and fluently as possible.

5. Part 2/Conditions -

a. The four scenarios must be completed within 20 minutes.

b. The examinee is permitted to quickly read the instructions for the scenario, but the use of lexical aids is not permitted.

c. Vocabulary and grammatical features used in the stimulus must be limited to those covered in the course of instruction for which the examinee is being measured.

6. Part 2/Criterion -

a. Scoring is done as described in 3a and b above.

b. No criterion is prescribed for Part 2, but a 63 raw-score cut-off (equivalent to an S-3 Level) is established for the entire Speaking test.

Appendix D

Spanish MCRT Net Administration Time

Skill Measured	MCRT Administration Time (in minutes)			TOTAL
	ANCHOR CRT #1	ANCHOR CRT #2	FINAL EXAM	
Listening Comprehension	25	25	30	80
Reading Comprehension	15	20	50	85
Translation	15	30	45	90
Speaking*	15	20	30	65
Writing	20	45	60	125
Number Transcription	15	15	15	45
General Transcription	60	90	135	285
TOTAL	165	245	365	775

* With the exception of the Speaking Test, knowledge of the foreign language is not required for MCRT administration.

Appendix E

SPANISH SPEAKING ANCHOR CRT #2

STUDENT RATING SHEET

NAME ______________________________ DATE ____________________

SSN ______________________ CLASS NO. ________________________

S-2 COMPUTATION TABLE				
RATING LEVEL:	1	2	3	SKILL POINTS
LINGUISTIC CATEGORIES:				
Pronunciation	2	3	4	______
Vocabulary	8	10	12	______
Grammar	12	14	16	______
Fluency	4	5	6	______
Comprehension	10	12	14	______
			SCORE =	

CONVERSION TABLE 2-A SCORE TO LEVEL		
SCORE	=	LEVEL
Minimum Score 36	=	1
37 - 44	=	1+
45 - 52	=	2

CONVERSION TABLE 2-B SCORE TO PERCENTAGE			
SCORE	% SCORE	SCORE	% SCORE
52 =	100	44 =	69
51 =	98	43 =	65
50 =	94	42 =	61
49 =	90	41 =	57
48 =	85	40 =	53
47 =	80	39 =	49
46 =	75	38 =	45
45 =	70	37 =	41
		36 =	37

PASS ______________ SKILL LEVEL ______________

FAIL ______________ PERCENTAGE SCORE ______________

EXAMINER ______________________________________

Appendix F

ADDENDUM TO EXAMINEE'S GUIDE

SPANISH MAJOR CRITERION-REFERENCED ANCHOR TEST #2

SPEAKING

SPANISH BASIC COURSE
(Modules 4-6)

Performance Criteria

1. The Anchor #2 Speaking Test is designed to permit an accurate appraisal of your oral competency in Spanish when you have completed Module 6 of the Spanish Basic Course. The information provided here is to be used with the instructions provided in the Examinee's Guide for the Anchor #2 Speaking Test.

2. Your examiner will be a native speaker of Spanish who has been specially trained in the face-to-face oral interview technique. The examiner will base his/her judgment of your performance upon the linguistic quality of what you say during the interview.

3. Five linguistic categories have been identified as important to the oral communication process. The descriptive criteria which the examiner will use to judge your performance on the speaking test are presented on the following page. Each category has been subdivided into three parts and assigned a rating scale -- 1, 2, or 3. The rating scale for each category will determine the number of points you will receive on the test.

4. Certain linguistic categories are deemed to be of greater importance than others for speaking. Therefore, different weights have been assigned which reflect the relative priority of each linguistic category. As you can see from the Student Rating Sheet (last page in your Examinee's Guide), the priorities are, in descending order of importance:

Grammar
Comprehension
Vocabulary
Fluency
Pronunciation

Speaking Rating Scale for Spanish Anchor CRT #2

Category	Rating	Criteria
Pronunciation	3	An obvious foreign accent with occasional mispronunciations that cause misunderstanding.
	2	A marked foreign accent which requires concentrated listening, and mispronunciations which lead to frequent misunderstanding.
	1	Frequent errors and a very heavy accent make understanding difficult; requires frequent repetition.
Vocabulary	3	General vocabulary permits discussion of most topics listed, with some paraphrasing and circumlocutions.
	2	Choice of words frequently inaccurate, limitations of vocabulary prevent adequate discussion of some topics and situations.
	1	Vocabulary limited to a very basic level on the topics covered in the interview.
Grammar	3	Occasional errors, showing imperfect control of some major patterns, but seldom causing misunderstanding.
	2	Frequent errors, showing some major patterns uncontrolled and causing occasional irritation and misunderstanding.
	1	Constant errors, showing control of few major patterns and causing occasional irritation and misunderstanding.
Fluency	3	Speech is occasionally hesitant, with some unevenness caused by rephrasing and groping for words.
	2	Speech is frequently hesitant and jerky; sentence may be left urcompleted.
	1	Speech is very slow and uneven, except for routine phrases and social expressions.
Comprehension	3	Understands normal educated speech quite well, but requires occasional repetition or rephrasing.
	2	Understands careful, somewhat simplified speech, with considerable repetition, and rephrasing.
	1	Understands only slow, simple speech; requires frequent repetition and rephrasing.

References

Defense Language Institute. *Administration and Scoring Manual for Foreign Language Oral Production* (FSI Interview). Presidio of Monterey, Calif., 1965.

____________. *Spanish Basic Course*. Instructional Guide and Modules 1-9, Tests and Workbooks, Presidio of Monterey, Calif., 1975.

____________. *Spanish Basic Course*. Major Criterion-Referenced Tests (Examiner's and Examinee's Guides and Administration and Scoring Manual), Presidio of Monterey, Calif., Foreign Language Center, July 1976, August 1976, October 1976, November 1977.

____________. Systems Development Agency (Provisional), *Test Development Standards*. Presidio of Monterey, Calif., July 1974.

Department of the Army, *Interservice Procedures for Instructional Systems Development*. Fort Monroe, Va.: Headquarters, United States Army Training and Doctrine Command, August 1975.

Lowe, Pardee, Jr. *Handbook on Question Types and Their Use in LLC Oral Proficiency Tests* (Preliminary Version). Arlington, Va.: Language Learning Center, Central Intelligence Agency, May 1976.

Woodford, Protase E. "Testing Guidelines for DLI Tests." Princeton, N.J.: Educational Testing Service, 1972.

Additional references as listed in "TRADOC Pamphlet 350-30," *Interservice Procedures for Instructional Systems Development, Executive Summary and Model*, pp. 132-49. Fort Monroe, Va.: Headquarters, United States Army Training and Doctrine Command, August 1, 1974.

ORAL PROFICIENCY TESTING IN NEW JERSEY BILINGUAL AND ENGLISH AS A SECOND LANGUAGE TEACHER CERTIFICATION

Richard W. Brown

New Jersey State Department of Education

ORAL PROFICIENCY TESTING IN NEW JERSEY BILINGUAL AND ENGLISH AS A SECOND LANGUAGE TEACHER CERTIFICATION

Richard W. Brown

On January 8, 1975, New Jersey's governor, Brendan T. Byrne, signed Senate Bill No. 811, also known as the New Jersey Bilingual Law. The law provided for mandatory bilingual education programs in New Jersey public schools.

Regulations for use in administering programs in bilingual education require that teachers of bilingual and English as a second language education possess appropriate certification.

The New Jersey State Board of Education, on October 1, 1975, approved bilingual/bicultural and English as a second language teacher certification regulations. The State Department of Education's Bureau of Teacher Education and Academic Credentials maintains responsibility for monitoring the implementation of the regulations.

Bilingual/bicultural and English as a second language certification regulations were developed by a statewide committee of experts in bilingual and English as a second language education. The committee consisted of public school teachers, college and university staff, Department of Education staff, Educational Testing Service staff, and members of statewide bilingual interest groups. Prior to their final approval by the State Board of Education, the certification regulations underwent numerous revisions after having been reviewed by educational personnel throughout the state. The final draft of the regulations also appeared in the New Jersey State Register on two occasions.

English as a second language certification regulations require that all teachers display "evidence of native or near-native competency in English as determined by guidelines . . . established by the New Jersey State Department of Education." To be eligible for standard or substandard bilingual/bicultural certification, all teachers must provide "demonstration of verbal and written proficiency in English and in one other language used also as a medium of instruction."

Prior to the enactment of the certification regulations in 1975, the State Department of Education sought the assistance of Educational Testing Service to develop a method and/or device capable of determining (1) native or near-native competency in English and (2) proficiency in English and other languages used as media of instruction.

Teachers in bilingual and English as a second language programs are expected to possess sufficient language competency to adequately present subject matter and to conduct classroom activities.

According to Educational Testing Service staff, heretofore most measures of second- or foreign-language ability were designed primarily to assess those skills normally stressed in formal, academic foreign language programs. These measures were not well suited to determine the ability of the examinee to function effectively in the other language milieu. Emphasis in such tests was often on formal grammar, grammatical terminology, and literary analysis--areas of questionable need for many bilingual teachers.

The need, therefore, was for an examination or a procedure that would measure the ability of the examinee to function effectively in the classroom through the medium of English (for teachers of English as a second language education) or English and Spanish (for teachers of bilingual education). The ability to function effectively would be manifested by such things as (1) the ability to comprehend completely the "talk" of children and parents, both English speaking and Spanish speaking; (2) the ability to communicate in English and Spanish with children and parents on school-related and other topics; and (3) the ability to present subject matter in the classroom, carry on classroom discussion, ask and answer questions, and explain concepts in both English and Spanish.

An issue of importance equal to that of the measurement of language proficiency is the determination of minimum competency. That a bilingual teacher must be "fluent" in English and Spanish seems a reasonable qualification, but what does "fluent" mean? What level of language performance should be the requisite minimum for teachers to carry out their duties in bilingual classrooms?

The instrument and procedures developed by Educational Testing Service addressed two broad issues: (1) the evaluation of oral proficiency in English and Spanish and (2) the establishment of criteria for determining minimal competency in English and Spanish.

The system developed for the New Jersey State Department of Education by Educational Testing Service for the purpose of determining oral language proficiency in English and Spanish is known as the Language Proficiency Program.

The program utilizes the Language Proficiency Interview (LPI), which was developed by linguists at the Foreign Service Institute. The Foreign Service Institute provides foreign language training to and certifies the foreign language abilities of U.S. Department of State and other federal government personnel.

mong the reasons for the development of the Language Proficiency Interview procedure was the absence of a reliable, direct measure of communicative competence (listening comprehension and speaking skills) that would be appropriate to assess skills from the level of no ability to the level of proficiency equivalent to that of an educated native speaker.

The Language Proficiency Interview has been in use for over fifteen years. Among the federal agencies using the LPI and the accompanying scale are the Department of State, Department of Defense, and ACTION/Peace Corps.

The interview procedure as carried out by the Foreign Service Institute, the Peace Corps, and others is as follows:

The interviewee, the interviewer, and a rater/linguist meet for up to thirty minutes. During this period the interviewer carries out what appears to be a friendly, informal conversation with the examinee. The rater/linguist may join in the conversation when and if appropriate. The interviewer conducts the conversation in such a way that a relatively complete sample of the examinee's abilities in the target language is obtained. Typically, the interview begins at a relatively simple level and becomes progressively more complex. The vocabulary, structure, and comprehension required to continue the conversation become increasingly difficult. When the interviewer and rater/linguist are confident the examinee has performed at the highest level of which he or she is capable, the interview is concluded.

The length of the interview is usually in direct proportion to the ability of the examinee--i.e., the lower the level, the shorter the interview; the higher the level, the longer the interview. The normal extremes are ten and thirty minutes.

Although it is common for the interviewer and rater to confer and agree on a rating, the responsibility for the official rating rests with the rater/linguist.

In addition to the conversation per se, one or more activities designed to elicit further evidence of the examinee's ability may be undertaken, such as a series of direct translations or a "real-life" situation in which the examinee serves as interpreter between a "monolingual English" and a "monolingual Spanish" speaker.

All applicants for New Jersey bilingual/bicultural and English as a second language certification must complete Language Proficiency Interviews. An applicant seeking bilingual/bicultural certification must complete Language Proficiency Interviews in English and the other language he or she will use in the public school classroom as the medium of instruction. An English as a second language certification applicant must complete an LPI in English.

In New Jersey, Language Proficiency Interviews may be completed at any one of seven centers established by the State Department of Education with the assistance of Educational Testing Service. The centers are located at Glassboro State College, Jersey City State College, Kean College of New Jersey, Monmouth College, Rutgers Graduate School of Education, Trenton State College, and William Paterson College of New Jersey.

The State Department of Education utilized two principal criteria when determining sites for centers: each had to be (1) an institution of higher learning offering a bilingual and/or English as a second language teacher education program and (2) located near public school districts containing large populations of bilingual students and teachers.

Interviewers for the centers were identified, screened, and selected for training by the State Department of Education with the assistance of Educational Testing Service. The trainees were language specialists from New Jersey public schools and institutions of higher learning. All trainees participated in training sessions conducted by Educational Testing Service. Upon completion of the sessions, the participants were certified as official language proficiency interviewers if they met all qualifications identified by Educational Testing Service, including the ability to reach an oral language proficiency level of 4 in the languages in which they were trained to interview.

As of March 1, 1978, applicants for English as a second language certification must reach a proficiency level of 4 in English to be eligible for standard certification. A level of 3 in English and 4 in the other language used as the medium of instruction are required for standard bilingual/bicultural certification.

To date, more than 1,400 Language Proficiency Interviews required for New Jersey bilingual/bicultural and English as a second language teacher certifiction have been completed.

During the past two years I have been asked, on a number of occasions, what I consider to be the strengths of the New Jersey program, and what recommendations I would give to any state planning to develop certification in these areas.

I will first list what I consider to be the strengths of our program:

1. Certification regulations were developed by a statewide committee of experts in bilingual and English as a second language education, including a representative of the state education association.

2. Certification regulations require language proficiency for both certificates.

3. Educational Testing Service has been assisting New Jersey from the beginning in the development of the certification regulations and the language proficiency interview system.

4. Oral language proficiency for teachers is determined by use of the Foreign Service Institute language proficiency interview and scale.

5. Language proficiency interviews are given in a number of regional centers strategically located throughout the state so as to provide teachers easy access to centers for interviews.

6. Interviewers are trained by Educational Testing Service.

7. The high levels of proficiency required for certification assure greater opportunities for successful communication between teachers and students in the classroom.

8. The comprehensive certification regulations guarantee that all teachers possess appropriate background needed to be more effective in the classroom. Regulations for both certificates contain extensive cultural components. The English as a second language regulations provide for comprehensive study in linguistics.

9. The results of recent litigation regarding the certification regulations have strengthened the overall program. Federal and state courts have determined that the regulations are legal, fair, and nondiscriminatory.

Second, I will identify some suggestions I would give to states planning to develop certification regulations for bilingual and English as a second language teachers:

1. Provide for funding at the state level to support the implementation of bilingual legislation.

2. Communicate with state legislators during the developmental stages of legislation.

3. Involve representatives of all statewide interest groups, including public school teachers and administrators, when developing regulations.

4. Require oral language proficiency in English for teachers of English as a second language, and in English and the other language being used as the medium of instruction in the classroom for bilingual teachers.

5. Utilize the Foreign Service Institute language proficiency interview system.

6. Request the assistance of Educational Testing Service when developing an interviewing system.

7. If possible, pretest the language proficiency system chosen for state use prior to implementing such a program. This should include conducting validity and reliability studies.

8. Require that tapes of interviewees be rated by more than one rater.

9. Contact other states that have instituted regulations to request information regarding their developmental and implementation procedures.

10. Develop regional interview centers within the state, as has been done in New Jersey.

11. Train prospective interviewers who have appropriate bilingual and/or English as a second language educational experience.

12. Work closely with institutions of higher learning that wish to develop teacher training programs.

13. Consider all areas previously identified as strengths of the New Jersey program.

14. Provide discussion sessions throughout the state for teachers who will be affected by regulations. At that time, explain all ramifications of the implementation of the regulations, including the language proficiency interviewing and rating systems.

15. Educate the public. Provide information to parents of children who will be affected by the regulations, either through workshops or with printed materials.

16. Provide opportunities for teachers who possess teaching experience in bilingual and/or English as a second language classrooms to be given credit for such experience. The credit should be applicable toward standard certification.

17. Provide all parties concerned sufficient time to fulfill all rules and regulations related to bilingual and English as a second language certification.

References

An Approach to the Assessment of English and Spanish Oral Proficiency of Bilingual/Bicultural Teachers and Teacher Candidates. Princeton, N.J.: Educational Testing Service, June 1976.

Language Proficiency Program. Bulletin of information. Princeton, N.J.: Educational Testing Service, 1976.

New Jersey State Department of Education, Bureau of Bilingual Education. *New Jersey Bilingual Law*. Trenton, January 8, 1975.

____________. *Regulations for Use in Administering Programs in Bilingual Education as Provided for in Chapter 197 of the New Jersey Laws of 1974*. Trenton, 1975.

New Jersey State Department of Education, Bureau of Teacher Education and Academic Credentials. *New Jersey Bilingual/Bicultural Teacher Certification Regulations*. Trenton, October 1975.

____________. *New Jersey English as a Second Language Teacher Certification Regulations*. Trenton, October 1975.

ADAPTATION OF THE FSI INTERVIEW SCALE FOR SECONDARY SCHOOLS AND COLLEGES

Claus Reschke

University of Houston

ADAPTATION OF THE FSI INTERVIEW SCALE FOR SECONDARY SCHOOLS AND COLLEGES

Claus Reschke

The prototype for the direct oral interview proficiency tests currently in use by U.S. government agencies and in a few schools and colleges is the interview test developed in 1956 by the staff of the Foreign Service Institute (FSI) of the U.S. Department of State. Although this test has undergone several changes and refinements during the past twenty-two years, its original format is still basically intact. This is because the test has, over the years, repeatedly proven itself to be a highly face-valid, extremely reliable and--for the specific needs of the FSI--very practical vehicle with which to determine the oral proficiency of career diplomats and other foreign service personnel whose jobs require foreign language proficiency.[1]

Because this particular test meets so well the basic criteria of reliability and practicality, if not also the criterion of validity, an increasing number of educators teaching foreign languages in high schools and colleges are considering using it to determine the oral proficiency of their students at various points during their language study. High school teachers could use the test to measure the oral proficiency of their students after two, three, or four years of language study. In college the test could have several uses. It could, of course, measure the oral proficiency of students after two, three, or four semesters of language study. It could also serve as part of a diagnostic and qualifying examination in undergraduate foreign language education programs, to assure that only those students who have reached at least an oral proficiency level of 2 are allowed to start the student-teaching phase of their programs.[2] At the graduate level, the test could be used as part of a qualifying examination for admission to graduate programs and for awarding teaching fellowships in foreign language departments.

[1] Those unfamiliar with the FSI test can find a detailed description of it in the article "The Oral Interview" by Claudia P. Wilds, one of the originators of the test, in Testing Language Proficiency, edited by Randall L. Jones and Bernard Spolsky (Arlington, Va.: Center for Applied Linguistics, 1975), pp. 29-44.

[2] A very elaborate interview system is being used by Purdue University in it's teacher education program. There, each undergraduate major in teacher education must complete two interview sessions with a three-person testing team, consisting of the coordinator for foreign languages and literatures education, a methodologist in the target

One of the prime reasons why this test is of such interest to teachers who wish to assess the oral proficiency of their students is the test's high reliability. A cross-language reliability study, conducted by the FSI in 1973, included French, German, and Spanish tests, and yielded a reliability coefficient of .85. Other in-house reliability studies conducted by the FSI, which were limited to only one language, have produced similar results, with one study, based on French tests given, showing a reliability coefficient of .93.[3]

Another reason why this particular oral proficiency test is of great interest to high school and college teachers is the thorough evaluation criteria set up for it by the FSI. Table 1 shows that the FSI evaluates a candidate's interview performance in five categories: accent (pronunciation and intonation), grammar (morphology and syntax), vocabulary, fluency, and comprehension. A weighted point system has been developed by the FSI, with the weights distributed as follows: accent 0, grammar 3, vocabulary 2, fluency 1 and comprehension 2. Thus grammar, vocabulary, and comprehension are considered by the FSI to be the most important elements of oral proficiency, a view most language teachers would be able to support on the basis of their own experience. The FSI's weighted scoring system (Table 2) was derived from multiple-correlation studies using the level ratings that had been assigned to numerous examinees.[4]

language, and an instructor in the target language. The first interview is diagnostic in nature; the second one, given at the completion of an advanced conversation course in the target language, seeks to determine if the student meets predetermined minimal oral proficiency standards before he or she is given permission to start the semester of student teaching.

At the University of Houston, an interview test, conducted by three faculty members, is used only in the German teacher education program. It is part of a comprehensive examination on language, culture, and literature that every German teacher education major must pass before starting the semester of student teaching.

[3] For the results of a more recent reliability study of FSI test scores, see Marianne L. Adams's paper in this volume: "Measuring Foreign Language Speaking Proficiency: A Study of Agreement among Raters."

[4] Wilds, p. 32.

TABLE 1

FSI Speaking Evaluation

			1	2	3	4	5	6	
1.	Accent	foreign	4	3	2	2	1	0	native
2.	Grammar	inaccurate	6	12	18	24	30	36	accurate
3.	Vocabulary	inadequate	4	8	12	16	20	24	adequate
4.	Fluency	uneven	2	4	6	8	10	12	even
5.	Comprehension	incomplete	4	8	12	15	19	23	complete

TABLE 2

FSI Level Assignment

FSI Score	FSI Rating
0-15	S-0
16-25	S-0+
26-32	S-1
33-42	S-1+
43-52	S-2
53-62	S-2+
63-72	S-3
73-82	S-3+
83-92	S-4
93-99	S-4+

However, there are two major reasons why the FSI interview test, in its present form, is not really suitable for use in high school and college.

First, the test's administration, which has proven to be very practical for the FSI, would be much less practical for schools and colleges. As it stands, two testers are required for each testing session,[5] one a native speaker of the target language and the other a certified language examiner, who may be either a native speaker and instructor of the target language or a linguist thoroughly familiar with the language.[6] Past experience of the FSI, CIA, and Peace Corps has shown that an examination team is able to conduct about fifteen interviews per day.[7] Since schools and colleges must test hundreds of students at the end of a term or a semester, however, the man-hours involved would be almost prohibitive. In addition, administering the test costs an estimated $40 per examinee,[8] a figure that, when multiplied by hundreds of students, would also be prohibitive.

The second major problem with using the FSI test in high school and college lies in the absolute oral proficiency rating scale used by the FSI and other government agencies. Ranging from 0 to 5--that is, from almost no speaking ability to a thoroughly bilingual fluency, with a "plus" level above each primary level[9]--the scale is far too broad in scope to be meaningful for use when testing the limited oral proficiency found in high schools and colleges. John Carroll's well-documented study of 1967, which sought to determine the foreign language proficiency of college language majors, revealed that few of them ever

[5] Of the five government agencies administering the interview test (FSI, DLI, NSA, CIA, and CSC), only the Defense Language Institute uses, due to limited resources, one tester. See Pardee Lowe, Jr., The Oral Language Proficiency Test (Washington, D.C.: Interagency Language Round Table, 1976), p. 2.

[6] See Wilds, p. 30. Before a language examiner can be certified, he or she must have reached at least the oral proficiency level 4 in the target language.

[7] John L. D. Clark, "Theoretical and Technical Considerations in Oral Proficiency Testing," Testing Language Proficiency, p. 16.

[8] This figure is based on information supplied for the year 1977 by the Testing Committee of the Interagency Language Round Table, U.S. Government.

[9] The "plus" designation indicates that a candidate has reached a proficiency that substantially exceeds the minimum requirements for a given level but does not meet all the minimum requirements for the next higher level. See Wilds, p. 36.

reached the 2+ level on the FSI scale during their senior year, whether they were studying French, German, Russian, or Spanish.[10] I believe this situation has not changed much in the past ten years. Therefore, most of the students tested in high school and college would fall into only three FSI categories, 1, 1+, and 2, making it difficult to show differences among them or to indicate their progress over a period of one or two semesters.

It appears, therefore, that before the FSI interview test can be used effectively in high schools and colleges some major modifications are necessary.

Suggested Modifications to Interview Procedure and Scale

I believe that the excessively high time and cost factors related to the administration of the test could be reduced without much loss in the reliability of the test results. The method I suggest is to reduce the testing team from two to one and to increase the number of students tested from one at a time to three, four, or even five. I believe the test would then be practical and would also remain a reliable instrument, so long as care were taken that all students being tested at the same time were at about the same level of proficiency.

The second problem with using the FSI test in high school or college--the broad absolute proficiency rating scale--is more complex but also has a solution. The solution I propose is to modify the FSI rating scale. Let us add to the six whole numbers and the five "plus" levels used by the FSI a second series of numbers that will refine the examinee's score and make it more meaningful. Each FSI number can be followed by a decimal point, and then by one or more additional "fine-tune" or performance-interpretive numbers.

I see this proposal as a combination of two scales, one vertical and one horizontal. The FSI ratings fall on a vertical scale:

0+
1
1+
2
2+
etc.

[10] John B. Carroll, Foreign Language Attainments of Language Majors in the Senior Year: A Survey Conducted in U.S. Colleges and Universities (Cambridge, Mass.: Harvard University, 1967), pp. 10 ff., 40 ff.; John B. Carroll, "Foreign Language Proficiency Levels Attained by Language Majors Near Graduation from College," Foreign Language Annals I, No. 2 (1967), pp. 131-51.

To this scale I would add a scale of horizontal numbers at each of the vertical scale levels, designed to provide as much precise data about a student's linguistic performance as a teacher might want.

For example, two students' oral proficiency may lie somewhere between the FSI ratings of 0+ and 1. Which of the two students is more proficient? The horizontal scale might indicate that the first one has a fine-tune score of 3 and the second a score of 7. The total ratings for these students could then be written as 0+.3 and 0+.7, visually awkward ratings to which I shall return shortly. The second student has, in any case, been shown to be more proficient--on the basis of the combined vertical and horizontal scales.

Theoretically, it would be possible to add an infinite number of digits to the horizontal scale. For example, the fine-tune digits 3 and 7 in the above example could be followed by five other digits indicating, on a scale of 0 to 9, the strength of the student's performance in each of the five evaluated categories (accent, grammar, vocabulary, fluency, and comprehension). Six additional digits might represent diagnostic ratings, with the first digit again a composite rating, on a scale of 0 to 9, followed by the five digits representing individual ratings in the five evaluated categories. These digits could, for example, provide information in the areas of phonology and syntax that would show whether a student has started to internalize a faulty phonological or grammatical system, and to what extent. Another group of six digits, the first one again a composite of the following five, could represent a specific projection of the degree of success that might be expected from future language training in each of the five evaluated categories.

The possibilities for use of the horizontal scale seem endless. However, the value of expanding it beyond the composite rating for each of the three proposed major areas (fine-tune, diagnostic, and projection) is questionable, since detailed ratings in only these three areas would result in an overall rating nineteen digits long. This would be an extremely awkward number to read and interpret. Retaining only the composite rating digit for each area, on the other hand, would yield a total rating for each test performance of only four digits. This number would certainly provide both student and teacher with far more information about the student's linguistic performance on the test than the single-digit FSI level assignment yields.

Of course, narrative descriptions would have to be written for each point on the horizontal scale at each of the eleven proficiency levels. The task seems enormous. It could be simplified, however, if only three narrative descriptions were written for each of the three areas (fine-tune, diagnostic, projection) proposed for the horizontal scale. Each area would then have a narrative description for the subranges 0-3, 4-6, and 7-9. Furthermore, since high school and college students would seldom exceed the 2+ level on the FSI absolute oral proficiency rating scale, why not limit the narrative descriptions for the horizontal scale to the 0+ to 2+ range on the vertical scale?

I recommend that the FSI rating scale be modified only in these ways, however, and not in others. I would retain the weighted scoring system used by the FSI and the present level assignment system, where the level is determined by the number of points achieved by the examinee in each of the five categories in which his performance is being rated (see Tables 1 and 2).[11] Both have proven over the past twenty years to be highly reliable measures of oral proficiency. I would suggest, however, that all eleven points on the FSI absolute oral proficiency rating scale be converted into two-digit numbers to facilitate recording of the test results. Thus level 1 would be recorded as level 10, level 1+ as 15, and level 0+ as 05.[12] This procedure would keep intact the narrative descriptions developed by the FSI for each general proficiency level and allow us to continue to indicate a strong test performance that warrants a plus rating without having an awkward plus sign preceding the decimal point. Also, the chance of an administrative error occurring in the recording of the student's rating on his permanent school record would be substantially reduced by changing the plus sign to a number, an aspect not to be treated lightly in this period of increased reliance on computerized record-keeping systems in high schools and colleges.

Example of Expanded Diagnostic Scale

So far I have discussed the possibilities of adapting a few administrative procedures and the rating scale of the FSI interview test to meet the realities and needs of high school and college teachers. I would like to concentrate on only one of the three areas on the proposed horizontal scale, the one that involves the first digit after the decimal point. This is the most important of the three digits, because it contains the most useful information for teacher and student alike: the progress a student has made during a given period of time--say, one or two semesters.

[11] However, I would suggest that the range of points in the first category on the FSI scale, accent, be reversed, since it makes little sense to award zero points for a "native" accent and four points for an obviously "foreign" one. The number of points involved is nominal.

[12] It may be argued that the conversion of the "+" to the digit "5" creates a false impression, since the FSI assigns a plus rating only to a performance that substantially exceeds the minimum requirements for a given level but does not meet all the minimum requirements for the next higher level. Use of the digit "5" to indicate a plus rating seems to imply, however, that the candidate's linguistic performance (on a scale of 0-9) met half the minimum requirements for the next higher level, not most of them, as FSI criteria demand. (See Wilds, p. 36.) The objection is valid, the problem minor. All that is needed is to substitute for the "5" a "7" or an "8" to convert the "+" to a numeral.

This first composite digit after the decimal point designates the fine-tune level of an examinee's linguistic performance. For this first digit on the horizontal scale, I propose the following preliminary narrative descriptions. They have been written using as a guide "Descriptions of the FSI Absolute Oral Proficiency Rating Scale" and the "Detailed Description of the FSI Checklist"[13] developed by the FSI in 1961.

Fine-Tune Level Description

General proficiency level: 05

Range 05.0-05.3: Candidate's pronunciation is nearly unintelligible; his use of grammar is almost always inaccurate; his vocabulary consists mostly of isolated high-frequency words that he uses haltingly; his ability to converse is extremely limited and does not go beyond answering simple yes/no questions.

Range 05.4-05.6: Candidate's pronunciation is frequently unintelligible; his use of grammar is often incorrect; his vocabulary is extremely limited and insufficient to carry on even the most simple conversation; his speech is halting and consists of individual words and simple phrases; his conversational skill barely goes beyond the ability to answer simple yes/no questions.

Range 05.7-05.9: Candidate's pronunciation is occasionally unintelligible; his use of grammar is frequently incorrect, preventing communication, but he shows some control over one or two major grammatical patterns; his vocabulary is quite limited, but he is able to carry on, though very haltingly, the most simple and fragmentary conversation about himself and his family (telling time, naming simple after-school activities, talking about main meals, telling the size of his family, and so on); he understands only slowly spoken speech and often-repeated simple statements and questions.

General proficiency level: 10

Range 10.0-10.3: Candidate frequently makes major pronunciation errors that impede understanding and require him to repeat his utterances; his rate of grammatical errors is extremely high, but he has some control over two or three major grammatical patterns, which he employs correctly with a fair degree of consistency, so that communication, although frequently hampered, is not entirely impossible; his range of vocabulary is

[13] Lowe, pp. 29-30.

limited to the basic personal and social level (e.g., time, three or four food items, two or three beverages, primary means of transportation, major weekend activities); his speech is slow and uneven; he understands very simple speech based on high-frequency situations or topics of a personal or social nature (e.g., age, simple family relationships, simple activities performed around the house, living accommodations at home), but requires frequent repetition and rephrasing of questions and statements.

Range 10.4-10.6: Candidate occasionally makes major pronunciation errors that interfere with understanding him consistently; his rate of grammatical errors is high, but he has good control over two or three major grammatical patterns, which he employs correctly with a high degree of consistency, allowing him to communicate at a fairly simple level; his vocabulary, although still limited to the basic personal and social level (e.g., four to ten food items, three to four beverages, simple purchases, the departure times of trains, planes, buses, and streetcars), allows him to communicate very briefly, simply, and imperfectly on a variety of high-frequency topics (e.g., daily meals, ordering two or three simple meals in a restaurant, describing in simple terms three to four activities at home, describing in simple language a visit to a grocery store, movie, theater, or concert, asking for simple directions); his speech is slow and uneven, except for short, routine sentences and phrases; his understanding is slow, although he does understand very simple statements and questions about a variety of high-frequency situations he would be expected to encounter daily, socially, or as a tourist, even though he may require frequent repetition and rephrasing of statements.

Range 10.7-10.9: Candidate seldom makes major pronunciation errors, but frequent minor errors hamper understanding; he makes many grammatical errors but has good control over three or four major grammatical patterns, which he employs correctly with a moderate degree of consistency, allowing communication to proceed at a fairly simple level; his vocabulary enables him to perform a variety of linguistic tasks (e.g., giving simple directions, asking for lodging, ordering fifteen to twenty-five different items of food and six different beverages, inquiring about the cost of postage, purchasing some items of clothing), even though his choice of words is frequently inaccurate; his speech is hesitant, and his sentences are very often left incomplete; he understands slow, simplified speech on a variety of personal, social, and tourist topics, but requires frequent repetition.

General proficiency level: 15

Range 15.0-15.3: Candidate occasionally makes minor pronunciation errors and has a distinctly foreign accent, which requires highly concentrated listening and leads occasionally to misunderstandings; his grammatical errors are of such a nature as to indicate that there are three or four grammatical patterns over which he has no consistent control (e.g., auxiliary verbs in perfect tenses, past participles of verbs, word order), causing occasional irritation and leading frequently to misunderstandings; he sometimes chooses incorrect words, but his vocabulary is large enough for him to be able to converse haltingly about routine travel needs (e.g., changing money, asking for and giving simple directions, ordering three different major meals, making simple introductions, making simple telephone calls, planning a trip with a travel agent) and a select group of topics in the personal and social domain (e.g., family, hometown, education, occupation or planned career); he understands quite well careful, somewhat simplified speech, but requires occasional repetition and rephrasing of statements.

Range 15.4-15.6: Candidate makes few pronunciation errors but has a strong foreign accent that requires concentrated listening; his grammatical errors are consistent enough to be categorized; his range of vocabulary allows him to talk with confidence about himself and other people, make introductions, discuss in simple language major events, describe medical needs to a nurse or pharmacist in simple terms, arrange a meeting with someone, and communicate to a service station attendant routine maintenance instructions for his car; his speech is sometimes jerky, often hesitant; occasionally sentences may be left uncompleted; however, he understands quite well somewhat below normal-rate speech that has been slightly simplified for his benefit, although some repetition and rephrasing of statements is required.

Range 15.7-15.9: Candidate's accent is quite foreign sounding and requires some concentrated listening; his pronunciation errors are few and mostly random; grammatical errors are of two kinds, random and consistent (some grammatical patterns are used incorrectly); his vocabulary range allows him to discuss in simple language, using many circumlocutions, some current events and a few high-frequency situations and topics of his own or his father's profession; his speech is hesitant; he frequently gropes for words and may need two or three starts before completing a sentence; he understands fairly well normal-rate, but somewhat simplified, speech; however, he may require the speaker to repeat or rephrase a comment occasionally.

General proficiency level: 20

Range 20.0-20.3: Candidate's accent is markedly foreign; he makes few but consistent pronunciation errors; his grammatical errors, which occasionally lead to misunderstandings, show that he lacks complete control of some major grammatical patterns; his range of vocabulary is adequate to handle confidently but not fluently inquiries and casual conversations about family and friends, current employment, trips, and his studies, using simple constructions and circumlocutions; his speech is somewhat hesitant; at times he gropes for words; he comprehends normal-rate speech quite well, only occasionally asking for the repetition of a word or phrase.

Range 20.4-20.6: Although the candidate's accent is foreign, his few mispronunciations are mostly random and only occasionally interfere with understanding; his infrequent grammatical errors show imperfect control of several grammatical patterns, but they seldom lead to misunderstandings; his vocabulary allows him to express himself, using simple constructions, quite accurately and with some confidence on a number of topics, including current events as well as his daily routine, studies, work, hobbies, and interests; he is able to describe a person or place in some detail, can narrate a sequence of events, and can ask in simple language for help when he sees himself confronted with difficulties or complications in his studies or his work; his speech is confident and only occasionally interrupted by groping for words; his comprehension of normal, educated speech is not perfect and requires the speaker occasionally to repeat or rephrase his sentences more simply.

Range 20.7-20.9: Candidate's few mispronunciations are slight and random; his accent is foreign; neither shortcoming seriously interferes with understanding; most of his grammatical errors are also random and seldom interfere with understanding; his vocabulary is sufficiently large that he can express himself simply and with some circumlocutions on a few social and professional topics, as long as they are general enough in nature not to require specialized vocabulary; his speech is somewhat uneven, caused by occasional rephrasings of sentences; his comprehension of normal, educated speech is nearly perfect, and he rarely requires sentences to be repeated or rephrased.

General proficiency level: 25

Range 25.0-25.3: Candidate's accent, although foreign, and his mispronunciations, which are minor and random, rarely lead to misunderstandings; random grammatical errors are frequent; consistent grammatical errors that show imperfect control of grammatical patterns are limited to two or three; his choice of words is sometimes inaccurate, but his vocabulary range permits him to discuss with some difficulty general student, professional, and social problems (e.g., financial problems, car repair, house repair/rebuilding, health problems); his speech is occasionally hesitant, caused by groping for the correct word; he understands normal, educated speech and seldom needs to have statements rephrased or restated for him.

Range 25.4-25.6: Candidate's accent is recognizably foreign; his errors in pronunciation are frequent but of little consequence with regard to understanding; occasional grammatical errors are random; one or two imperfectly controlled grammatical patterns lead to consistent errors, which, however, have little effect on understanding; his vocabulary includes a number of professional terms that extend the range of professional topics he is able to talk about; his speech when talking about more specialized professional topics is hesitant and marked by frequent groping for the correct words, but he comprehends most conversations of a nontechnical nature and some of a specialized, professional one.

Range 25.7-25.9: Although candidate's accent can still be classified as foreign, his rare errors in pronunciation do not interfere with communication; his grammatical errors are few, mostly random, except for perhaps one recurring pattern of error; his vocabulary inventory is large enough to allow him to discuss some special, professional interests with a colleague, although he uses simple constructions and interrupts his speech frequently to grope for the correct word; consequently, his speech is somewhat uneven, but he understands a native speaker of the target language well, except for very colloquial or too technical speech.

There is no need to reinvent the wheel. The FSI interview test is in principle the best oral proficiency test we have. Its reliability is high; its administration and evaluation procedures have been developed, tested, and retested numerous times over the past two decades by government testing teams. These factors are invaluable to those educators who seek to find a testing instrument with which to measure accurately the oral proficiency of their students.

I believe the few minor changes I have suggested in the test's administration procedure, and the major adaptation I propose here for its rating scale, meet the two basic objections frequently leveled against the FSI test when its use outside the government is being debated: the excessive amo [illegible] of time and money required to administer it, and the too broadly conc [illegible] FSI proficiency levels, which are not very meaningful when testing the limited oral proficiency of high school and college students.

INTERVIEW TECHNIQUES AND SCORING CRITERIA AT THE HIGHER PROFICIENCY LEVELS

Randall L. Jones

Brigham Young University

INTERVIEW TECHNIQUES AND SCORING CRITERIA AT THE HIGHER PROFICIENCY LEVELS

Randall L. Jones

Despite its acknowledged shortcomings, the oral interview remains the most useful and valid instrument for measuring spoken language proficiency. It closely approximates a real language situation and provides a wide variety of speech samples for evaluation. It is also sensitive to the entire range of language proficiency, i.e., from 0 to 5 on the FSI scale. It is not calibrated finely enough to discriminate well within levels, but that, after all, is not its original purpose.

In 1973 I spent several weeks interviewing language testers at the CIA and the FSI. Among other things, I asked them what they felt were significant problems with the oral interview technique. One of the most common responses was that the higher proficiency levels were very difficult to evaluate. (The higher levels are to be understood here as 3+ and above.) There is little problem for a trained tester to discriminate between a 1+ and a 2, but there is less certainty when it gets into the area from 3+ to 5. It generally takes longer to administer an oral interview to an examinee whose proficiency is at a high level, but the problem is really more than a function of time.

I would like to suggest four principal reasons for the difficulty in evaluating oral proficiency at the higher levels. (1) The definitions for levels 4 and 5 are not specific enough to provide a basis for making a valid judgment. (2) The standard list of performance factors--grammar, vocabulary, fluency, pronunciation, and comprehension--is not sufficient to distinguish proficiency at the higher levels. (3) The nature of the oral interview is such that it does not provide an efficient method of eliciting language performance at the higher levels. (4) Because the number of examinees at the higher levels is relatively small, testers do not have the opportunity to develop a feeling for the important distinctions between and among these levels.

The matter of the proficiency definitions, I feel, is important, and the government language community should consider the possibility of making revisions. Levels 1, 2, and 3 correspond to natural stages of proficiency development, and the definitions capture these stages quite well. Level 1, for example, is often referred to as the "survival" level; i.e., the speaker can communicate in the language sufficiently well to take care of his important needs. But he has difficulty holding up his end of a conversation for very long, and his control of grammar and breadth of vocabulary are weak. Level 2 is often referred to as the "courtesy" level; i.e., the speaker is able to engage in sustained conversation without a great deal of effort, even though he may make numerous errors and may not be able to express himself precisely in many areas. He is confined more to <u>what</u>, <u>when</u>, <u>who</u>, and <u>where</u>, having difficulty with <u>how</u> and <u>why</u>. The 3 level speaker has, in a sense, "arrived." He has confidence in using the language, and he understands his own strengths and limitations. His ability for expression is very good in his own area of interest and fair to good in other general areas.

The definition for level 4, howrver, does not provide much help in making a satisfactory distinction between levels 3 and 4. The level 4 definition does introduce two new tasks: ability to "respond appropriately even in unfamiliar situations" and to "handle informal interpreting from and into the language." But these descriptions are very vague and nothing is said about what the unfamiliar situtions or interpreting task might ~~be.~~ One sentence in the definition for level 4 is especially troubling. It states that the level 4 speaker "would rarely be taken for a native speaker." My experience with German is that nonnatives are often told that they "speak just like a native German." Even a level 1 speaker can pass for a native if his pronunciation is good and he keeps his sentences restricted to those he can say without errors.

The definition for level 5 seems at first to be somewhat more satisfying in that it is the highest mark on the scale, the ultimate. The speaker's proficiency must be equivalent to that of an educated native speaker. The obvious question here is, how does an educated native speaker speak? What exactly is the absolute criterion against which we are judging all our examinees? Do we really have a good intuitive feeling about it?

The second reason mentioned above concerns the list of performance factors. There is no question that a level 4 speaker has better control over structure, vocabulary, etc., than the level 3 speaker, but I feel there is an additional factor that becomes important at this point: the sociolinguistic factor. I do not mean sociolinguistics in the broad sense, but rather those aspects of language that have more to do with social interaction than with imparting information. Common examples include expressing gratitude, responding to an expression of gratitude, excusing oneself, responding to such an excuse, expressing greetings and farewells, paying a compliment, receiving a compliment, declining an invitation, expressing surprise or annoyance or anger, complaining, and so on. Social communication also includes the use of hesitation words and other noncommunicative words and phrases. In many cases it does not concern what is said so much as when and how it is said. For example, in our own culture the proper response to a compliment is usually "thank you," but in many cultures that would be considered impolite. If we sneeze it is expected of us to say "excuse me," but in some cultures nothing is said, because it is not considered polite to draw attention to the sneeze. The beginner does learn standard phrases for expressing gratitude, excusing himself, or whatever, but the presumed standard phrases often found in the textbooks are in many cases seldom used by real native speakers. I suggest that sociolinguistic sensitivity be added to the list of performance factors, and that it be incorporated into the definitions for levels 4 and 5.

The oral interview is really not an interview in the strict sense of the word, but rather a conversation between two or more people. It is also a test in that one of the partners in the conversation is providing stimuli and the other one is giving responses. But there is a lot of room for variation, and the examinee can often avoid problem areas by talking around them. How can the examinee's "high degree of fluency and precision of vocabulary" really be demonstrated? The fact is that the interview technique is not notably efficient for eliciting specific speech samples beyond the 3 level. It requires a lot of time to obtain very little data. Other nonconversational techniques are thus necessary to get at the important aspects of proficiency at the higher levels. It is true that such techniques tend to be artificial and somewhat removed from real language situations, but they can nevertheless be valid indicators of language proficiency.

The fourth problem mentioned above relates to the fact that most testers are so rarely exposed to examinees above the 3 level that they do not develop a feeling for how 4 and 5 level speakers should perform. This also raises an interesting question: Is there really a need to test beyond the 3 level? I have heard the suggestion made that anyone who is obviously above the 3 level should be put into the category 3/5, that is, somewhere between 3 and 5. I do not believe there are any language-essential positions in the government designated at the 5 level, and probably very few at the 4 level. It seems that knowing a candidate is beyond 3 would be sufficient. This is, of course, a managerial and not a linguistic issue, but it seems that if there are five levels of proficiency, we have an obligation to develop suitable techniques for testing at each level. With regard to the training of testers, after the criteria for performance at the higher levels have been more clearly defined, samples of 3+, 4, 4+, and 5 level speakers can be recorded and annotated for training purposes.

I feel that at the present time the range of proficiency levels from 3+ to 5 is not properly understood. There is, however, good evidence that there are criteria that can distinguish among the specific levels within this large realm. In an attempt to get closer to the problem, I considered several methods of eliciting language performance from examinees that would be useful in evaluating the higher levels. The procedures are not new with me, and in some cases they have already been tried by oral interview testers. I ultimately decided on four techniques that I wanted to experiment with: (1) a picture-vocabulary task, (2) an anecdote retelling task, (3) a repetition task, and (4) a situation task. The language I chose for the experiment was German. Because the language performance of an educated native speaker is the ultimate criterion of judgment, I had five educated native speakers of German participate in the experiment, along with ten educated nonnative speakers. The four techniques are described briefly below, followed by a discussion of the results of the experiment.

Vocabulary is one of the five specified factors for evaluating performance in an oral interview, and there is no question that the breadth and precision of vocabulary increases as the language learner approaches the level of the native speaker. But it is often difficult to judge from an oral interview what words the examinee does and does not know. For this experiment I decided to select words that are quite low in frequency but broad in their range of occurrence, i.e., objects that are very much a part of everyday life but not often talked about. These are words that native speakers are certain to know but that nonnative speakers would be less likely to have learned. The stimuli were pictures from German magazines. (The objects are listed in Appendix A.) Subjects were shown the pictures one by one and asked to identify the specific objects by name. They were asked to say so if they did not know the word for a particular object.

For the retelling task, each subject read five short anecdotes in German and retold each one in his own words immediately after it was read. (See Appendix B.) He was allowed as much time as he wished to read each anecdote, but he was not allowed to refer to the printed version after he began to retell it. The anecdotes were quite short, so memory was not really an important factor.

For the repetition task, every subject listened to five recorded German sentences. (See Appendix C.) As each sentence was played the subject listened and then attempted to repeat it verbatim. The sentences ranged in length from three to five seconds, from ten to nineteen words, and from twenty to twenty-nine syllables. The idea for the task comes from a study done a few years ago by Merrill Swain and others at the Ontario Institute for Studies in Education. Swain rejects the notion that repetition or imitation is merely a perceptual-motor skill. She claims that if the utterance to be repeated is long enough (she used French sentences of about fifteen syllables), it has to be decoded, stored, recalled, and encoded. This task is, of course, impossible unless the subject has some degree of proficiency in the language. The higher the proficiency, the better the ability to process the sentence and repeat it. The hearer must somehow match the incoming signal against existing words and structures in the language that he has stored in his memory. If the words and structures are not there, the sentence--or at least part of it--will evaporate and he will not be able to repeat it successfully.

The fourth task was the elicitation of expressions in various situations in an attempt to get at some of the sociolinguistic elements of language proficiency. Each subject was given ten cards on which specific situtions were described. (See Appendix D.) He was asked to read each card and say how he would respond in the situation.

Of the five native speakers who served as subjects, two were undergraduate students at Cornell, two were graduate students, and one was the wife of a graduate student. All the nonnative subjects spoke English as a first language. One of them was an undergraduate student; the others were graduate students. All have lived in Germany for extensive periods, and it has been said of six of them (by people who are in a position to judge) that they "speak just like natives." Whatever the case, all of them would be rated 3+ or higher.

The picture-vocabulary test was administered first. It performed very well in distinguishing between the native and nonnative speakers, but it did not discriminate well among the nonnative speakers. Among the native speakers, three of the ten objects were identified using the same words, five were identified using various synonyms, and two were problematic because of the pictures. Among the nonnative speakers none identified the objects using the same word for all subjects; and for no object did all the subjects use an acceptable word. The number of objects correctly identified by the nonnative speakers ranged from zero to three.

The effectiveness of the picture-vocabulary task can be demonstrated by three of the objects: a ball of yarn, a calf (of a leg), and an earlobe. These objects, by the way, were the three that all the native speakers identified with the same word. None of the nonnative speakers knew the word for "ball of yarn," although several of them said "yarn." One knew the word for calf, and five knew the word for earlobe. There are numerous objects that can be used for this task, i.e., objects that are a common part of the culture but thai nonnative speakers learn very late in their acquisition of the language. I feel it is a good supplement to the oral interview for testing at the higher levels. It also seems possible to assign difficulty factors to the various objects for a specific language, thus assisting in making finer discriminations within the higher proficiency range.

The retelling task not only discriminated well between the native and nonnative groups, but it also distinguished among the members of the nonnative group quite well. In all cases the native speakers retold the anecdotes with all the essential facts and using all key vocabulary. The performance among the nonnative group was spread across a broad range. In a couple of cases, the point of the story was completely missed.

There were a couple of rather unexpected side benefits that made this task even more interesting. First, the native speakers tended to use a lot of little filler and transition words and phrases that were not in the original story; the nrnnative speakers did not do this. Second, in many cases the nonnative speakers used vocabulary from the original story, but incorrectly, e.g., used the wrong gender or an incorrect past tense form. And, finally, it was obvious that some nonnative speakers simply did not understand the meaning of some of the words. This affected the retelling of the story considerably. The retelling task was the most time-consuming of the four, but it was quite productive. I did not take the time to analyze each speaker carefully, but I am certain that the performance of the nonnative speakers could easily be rank-ordered according to specific observable criteria.

The repetition task was quick and very effective. All the native speakers performed well on this task, having little difficulty repeating the sentences without errors. The performance of the nonnative speakers, on the other hand, was once again spread across a wide spectrum. None of them performed as well as any of the native speakers, but one came very close. Problems related directly to the length of the sentence and the vocabulary in it. The less proficient nonnative speakers had difficulty completing some of the longer sentences and tendeu to omit unfamiliar words and phrases. Also, similar words in sentences caused some confusion. One sentence, for example, has the words Ausserdem and aussergewöhnlich. The similarity of the two tended to create some confusion. Again, I did not make a careful analysis of each performance, but I feel this task is an excellent technique for testing proficiency at the higher levels.

The situation task was, without question, the most disappointing, although I am not yet ready to give it up. Whereas the native speakers retold the anecdotes with enthusiasm, they responded to the situations rather unnaturally. In most cases, they had to think about them for a while. Two of the situations proved to be very unproductive: the "pretty shirt" and "being startled." Native and nonnative speakers alike seemed to be puzzled for answers. Some interesting observations were made during this task, although I am not certain how useful they would be for testing. When asking directions of the man on the street, most of the nonnative speakers began by saying "excuse me" (or the German equivalent), but none of the native speakers did. When responding to the salesman at the door, the native speakers merely said, "No, I'm too busy" or "I never buy anything at the door." Several of the nonnative speakers gave elaborate explanations. Although the task was less than successful in getting at the social communication I was looking for, I feel it can be developed into a useful technique, and further work should be done. Much depends on what the situation is and how it is described.

I feel these four techniques can be valuable in assisting the tester to make judgments at the higher proficiency levels. More research needs to be done to refine the techniques and to specify the criteria more closely. A bank of pictures, anecdotes, sentences for repetition, and situations can be built up, with each one tested and assigned a difficulty factor. It is hoped that the vague proficiency area between 3+ and 5 will thus be better understood and become easier to evaluate.

Appendix A

List of Vocabulary Items

(1) bottle cap (screw type), (2) calf (of a leg), (3) dog's nose, (4) dumbbell, (5) hubcap, (6) earlobe, (7) weather vane, (8) ball of yarn, (9) gasoline pump, (10) place mat.

Appendix B

Anecdotes

1. Moses Suppengrün in Krotoschin verdiente mit seinem Getreidehandel so viel, dass er seinen Sohn studieren lassen konnte. Zum erstenmal kam der junge Moritz von der Berliner Universität auf Ferien nach Krotoschin und sein Vater fragte ihn, was er nun eigentlich studiere.

"Philosophie", antwortete der Sohn.

"Wie heisst? Was ist Philosophie?"

"Will ich dir zeigen, was ist Philosophie.--Also de glaubst, de bist in Krotoschin, nicht wahr?"

"Ja, ich glaub', ich bin in Krotoschin", gab der Vater zu.

"Pass auf, werd' ich dir mit meiner Philosophie beweisen, dass de nischt bist in Krotoschin!"

"Nanu!"

"Also, wenn de bist in Krotoschin, dann bist de nischt in Posen?"

"Nein, dann bin ich nicht in Posen."

"Wenn de bist nischt in Posen, dann bist de doch anderswo?"

"Is richtig!"

"Nu, wenn de bist anderswo, dann bist de doch nischt in Krotoschin?"

"Is wirklich richtig", murmelt der Vater und verfällt in tiefes Nachdenken. Auf einmal gab er seinem Sohn eine gewaltige Ohrfeige.

"Was ist?" rief dieser. "Warum schlägst de mir?"

"Ich?" sagte der Vater und machte ein ebenso erstauntes Gesicht. Ich hab' dir nischt geschlagen! Wie kann ich dir schlagen, wenn de bist in Krotoschin und ich bin anderswo?"

2. Tünnes und Schäl sind gestorben. Der eine kommt in den Himmel, der andere in die Hölle. Eines Tages haben beide Urlaub, und sie treffen sich auf einer Wolke.

Der Schäl, der aus der Hölle kommt, erzählt:"Ach, wir arbeiten am Tage zwei Stunden, und das Quartier ist anständig und das Essen ist auch ziemlich gut."

Der Tünnes erzählt aus dem Himmel! "Wir müssen jeden Tag zwölf Stunden arbeiten!"

"Wie?" sagte der Schäl. "Wie kommt das denn?"

Tünnes: "Ja, wir haben eben zu wenig Leute!"

3. Es war kurz vor Weihnachten, als ein armer Bauernjunge an einem Fenster des Bürgermeisters eine fette Gans hängen sah. Er dachte: Mein liebes Gänschen, du hängst dort oben so einsam, ich will dich in eine gute Familie bringen.

Am Abend ging er heimlich mit einer Leiter zum Hause des Bürgermeisters. Langsam stieg er zum Fenster hinauf, an dem die Gans hing. Er hatte den fetten Vogel schon in der Hand, als er plötzlich die laute Stimme eines Polizisten hörte: "Halt! Was machst du dort oben?" Ohne die Nerven zu verlieren, antwortete der Junge: "Da bald Weihnachten ist, will ich dem Herrn Bürgermeister als kleine Überraschung eine fette Gans an das Fenster hängen." Der Polizist rief ärgerlich: "Unsinn, komm sofort herunter!" "Nun", meinte der Junge, "das ist wirklich schade, denn jetzt muss ich die Gans wieder nach Hause mitnehmen.

4. Ein junger Amerikaner, der wie viele in diesen Tagen im Sommer nach Europa gefahren ist, kommt auf seiner Reise auch nach Italien. In Rom kommt er in einem kleinen Restaurant beim Essen mit einem Italiener ins Gespräch. Man erzählte sich von den beiden Ländern, ihren Menschen und ihren Eigentümlichkeiten. Der Amerikaner will seinem Freund erklären, wie gross sein Land ist im Vergleich zu Italien oder anderen Ländern.

"Bei uns setzt man sich in einen Zug, und dann fährt man eine Stunde, mehrere Stunden, sogar einige Tage, und dann ist man immer noch in Amerika."

Da antwortet der Italiener unbeeindruckt: "Das kennen wir! Solche Züge haben wir bei uns auch."

5. Eine reizende junge Dame tritt in ein Seidenwarengeschäft. Der tadellos frisierte und geschniegelte Verkäufer überschüttet sie mit einer Flut von liebenswürdigen Redensarten, und da die junge Dame keineswegs prüde zu sein scheint, wird er immer verliebter.

"Was kostet dieses seidene Band?" fragte die hübsche Kundin.

"Einen Kuss der Meter!" antwortet schmachtend der junge Mann.

"Schön, packen Sie mir zehn Meter ein!"

Als dies geschehen war, sagt die junge Dame lächelnd: "Warten Sie, draussen vor dem Schaufenster steht meine Grossmama, die bezahlt für mich."

Appendix C

Texts of Repetition Sentences

1. Ausserdem werden in diesem Jahr aussergewöhnlich viele Studienräte in den Ruhestand treten.

2. Proteste gegen Kernkraftwerke hat es in den letzten Monaten in Hülle und Fülle gegeben.

3. Aber es geht mir heute Abend gar nicht um die Frage, ob die Stuttgarter Entscheidung richtig war oder nicht.

4. Die Sowjetunion hat viele Millionen Tonnen Getreide in den Vereinigten Staaten gekauft.

5. Gleichzeitig hat diese Meldung jedoch für die Schulen eine Schattenseite.

Appendix D

Situations*

1. You are looking for the tourist office in an unfamiliar city. You go to someone who is standing on the street to ask directions. You say . . .

2. You are a guest for dinner at someone's house. You have almost finished eating, and the hostess offers you more food. You would like some, and you say . . .

3. You are in a department store and you accidentally step on someone's foot. You say . . .

4. You are wearing a new shirt (blouse). Someone sees it and says, "That's really beautiful." You say . . .

5. You have been speaking with a friend for about fifteen minutes. You have an appointment now and must go. You say . . .

6. You are speaking with a friend. He (she) says something very startling about someone else. You say . . .

7. You are sitting quietly at a desk reading a book. Someone walks up and says something to you. You are startled because you did not hear him coming. You say . . .

8. You are invited to a party but you really do not want to go. You say (lie) . . .

9. You have been waiting for a friend for thirty minutes. Finally he (she) comes. You say . . .

10. The doorbell rings. You go to the door and find a salesman. He introduces himself and asks, "May I come in for a few minutes?" You say . . .

*For the experiment, the sentences were in German.

References

Jones, Randall L. "The FSI Oral Interview." In Advances in Language Testing, edited by Bernard Spolsky. Arlington, Va.: Center for Applied Linguistics, in press.

Swain, Merrill, G. Dumas, and N. Naiman. "Alternatives to Spontaneous Speech." [EDRS: ED 123 872]

Valette, Rebecca. Modern Language Testing. rev. ed. New York: Harcourt, Brace, Jovanovich, 1977.

TESTING SPEAKING PROFICIENCY THROUGH FUNCTIONAL DIALOGUES

I. F. Roos-Wijgh

Dutch National Institute for

Educational Measurement

TESTING SPEAKING PROFICIENCY THROUGH FUNCTIONAL DIALOGUES

I. F. Roos-Wijgh

In this presentation I will deal with the following: (1) the teaching of modern foreign languages in the Netherlands; (2) the function of CITO (Dutch National Institute for Educational Measurement), the institute where I work; (3) recent developments in the tuition of speaking proficiency; (4) the purpose of the CITO speaking proficiency tests and a description of the area of language behavior covered by these tests; (5) the form and function of the tests; and (6) expectations for the future.

Language Teaching in the Netherlands

Modern foreign languages play an important part in secondary education in the Netherlands. The reason for this is that our language is spoken by very few people in comparison with, for instance, the English language. Moreover, there are numerous contacts with the surrounding countries, in both the economic and the touristic spheres.

To give an impression of the smallness of the area of this part of Western Europe, the distance from Paris to Amsterdam is about the same as that from Boston to Washington. And in Paris they speak French, as you all know. Most of the Dutch population lives less than one hour's drive from Germany, where they speak German.

Consequently, there is in the Netherlands a great need for being proficient in at least one foreign language, and this is reflected in the curriculum of the secondary schools, in which about 30 percent (and often more) of the total time available is devoted to modern foreign languages. In the first three years of secondary education, English, German, and French are obligatory subjects; later on it is possible to drop one or two. You can also choose Spanish or Russian. Since the sixties the emphasis in language teaching has been more and more on the communicative aspect of language. One of the consequences is that now more attention is paid to speaking.

In sociolinguistics methods were developed for describing this communicative aspect of the language and these methods are the base of modern curriculum development of foreign language education.

CITO

The developments in foreign language teaching are reflected in the activities of the language department of CITO. This institute was established in 1968 by the Dutch government, with the object of promoting the development of objective tests for the educational field.

At first the language department occupied itself with the production of reading comprehension tests; later we also made listening comprehension tests (both for use as final examinations). Recently we developed criterion-referenced tests for the first years of foreign language teaching and have started a project to develop speaking proficiency tests. As this is a fairly new project, it is not yet possible to provide detailed information on the tests and their outcomes. But I will try to explain to you the underlying concepts and how the contents of the tests are determined.

Recent Developments

First of all I'll give you some more background information about recent developments in the tuition of speaking proficiency. In the present situation it is usually the teachers that decide how they will test speaking proficiency. This means in practice that they ask students to tell something about the literary works they have read, or put questions to them with reference to a text. In other words, the students are simply asked to "say something about something." The CITO project, "Testing of Speaking Proficiency," does not conform to this situation, but is based on new trends in the field of systems development in language learning.

Under the auspices of the Council of Europe, experts have defined a so-called threshold level. This level "may be seen as the lowest level of effective language use, thus defining a threshold at which language learning establishes general communicative ability minimally adequate to the general range of language situations in a speech community and which is thus an appropriate objective for initial language courses" (Council for Cultural Cooperation of the Council of Europe, 1973). It is essentially a level of oral communicative ability, designed for adult learners.

The model for the defini'ion of language-learning objectives specifies eight components, but I'll mention only the most important ones for our tests. They are (1) the situations in which foreign language will be used, including the topics that will be dealt with; (2) the language functions (or speech acts) the learner will fulfill (e.g., giving information, asking for information); and (3) the specific (topic-related) notions the learner will be able to handle. As noted above, this threshold level was essentially developed for adult learners. But now the author has published a special version of this model for foreign language teaching in schools (V.Ek, 1975). Some of the suggestions of this adapted version have already been realized in a number of schools. There are schools that have special one-week projects on, for instance, shopping. The first thing required of the students is no longer to say something about something but to say something in a given situation.

We are now working on tests that can serve as a sequel to this development. What we want to test is the ability to perform various speech acts in a foreign language in the form of a dialogue, with the student both taking the conversational initiative and responding. By "dialogue" is meant here the whole of the dialogues that take place in communicatively relevant situations in which one is confronted with persons using that language.

A description has been made of those situations that can be considered communicatively relevant when one is abroad or comes into contact with foreigners in one's own country. The language behavior in such a situation is specified by the situation itself and any parts thereof, the roles played by the speakers in the situation, the speech acts that have to be performed, and the specific informational aspects connected with the speech act in that situation.

Example

When you describe the situation	:	camping
a part of that situation may be	:	reception desk
the roles played are those of	:	receptionist/guest
speech acts to be performed (by guest)	:	asking for information giving information persuading refusing yielding expressing wishes expressing (dis)satisfaction
the specific informational aspects to be dealt with	:	site for the tent number of persons equipment quietness (at night) facilities time of arrival/departure

In this way we specified some fifteen situations, such as public transport, shopping, police station, entertainment, and camping. Of course there are numerous other situations; we only picked those that could be relevant to the majority of the learners. In these situations the theme of the conversation is intrinsically quite stereotyped. At a railway station you never ask, "What is the color of a return ticket today?"

There are also communicatively relevant topics that are not limited to particular situations, and the language behavior in these cases will be far less predictable. One can, for example, tell something about one's hobbies at the edge of a swimming pool, or at a party, or in the compartment of a train, etc. That is why a thematic specification of the language behavior required has been included in the description of the area of language behavior that will be covered by the tests. The theme has been further specified as follows: (1) the theme and its subthemes and (2) the speech acts to be performed with regard to the theme.

Example

When you pick out the theme: personal data

subthemes are : name, address, age, origin

and the speech acts to be performed can be : identifying
qualifying

We listed the following themes:

- everyday life
- holidays
- family/relatives
- personal education
- ambitions
- interests
- spending one's leisure time
- home
- hometown
- information on one's own country and people
- current social and political problems

The author of the threshold level concept does not make this difference between situations and themes; he just presents a list of topics. We, however, consider this distinction useful when you work out the system in more detail. "Railway station" can be a theme; you can talk about trains and railway stations anywhere. But when you consider it as a situation, i.e., when you take into account the setting of a railway station, you perform another kind of language behavior.

It is quite obvious that these descriptions are not exhaustive. Teachers will be consulted to find out what relevant themes and situations are still lacking. Moreover, they will have to indicate the priorities within the area of language behavior. Besides the speech acts that are linked to themes and situations we listed also a separate group of so-called social speech acts, which serve to start or end a conversation and to show courtesy, such as greeting, introducing oneself, inviting, thanking, taking one's leave, congratulating, and expressing best wishes.

Thus, the language behavior that is required by the tests can be classified according to situational specification, thematical specification, and social specification.

Form and Function

As the tests are based on a method of specifying language learning objectives that has only started to make its way into the schools, it would be premature to offer them as selective final tests now. We are developing them chiefly to support the learning process in schools. They will, therefore, be introduced in cooperation with other institutes rendering services to the field of education, such as the National Institute for Curriculum Development and regional school advisory centers. The test will be published in the form of a set of thematical and situational tests. An index will make it possible to choose several entries to the tests.

Use of the tests can best be illustrated with the help of a practical example. Suppose a teacher of French wants the students to be able to communicate their accommodation needs to the receptionist at a camping site. The teacher thus chooses from the index "situations" the test "camping." The test begins with a short introduction so the student knows what role he or she has to play.

The first tasks set in this test are:

1. Le soir, vous arrivez à la réception du camping. Là, il y a une vieille dame. Saluez la dame!

2. La dame dit "Bonsoir." Puis vous demandez une place à la dame.
   You can answer: Je veux/voudrais camper ici
   - une place (pour ma tente)
   - passer la nuit ici/au camping

   specification: role: guest/receptionist
   speech act: asking for information
   notion: site for tent

3. La dame vous dit: Il n'y a plus de place. Vous insistez, vous faites savoir que votre tente n'est que très petite.
   - (Mais Madame) (je vous en prie) ma tente est très petite.
   - même pas une toute petite place?
   - (Vous êtes sûre) même pour une toute petite tente?

   specification: role: guest/receptionist
   speech act: persuading
   notion: site for tent

And so on. (The test comprises ten tasks.) In the second part of the test you make the acquaintance of your neighbor at the camping site. The dialogue that follows can be characterized as a thematical dialogue; you are asked to talk about your country and your hometown. The conversation runs as follows:

Le voisin dit: Vous n'êtes pas français n'est-ce pas?

The student answers: Non, je suis Hollandais.

Le voisin: Ah, la Hollande. La capitale de Copenhague est magnifique!

The student answers: Copenhague n'est pas la capitale de la
Hollande. C'est Amsterdam.

When they were pretested, these items proved to work very well. The students got so involved that they simply forgot they were in a test situation and answered very spontaneously, even indignantly in the "persuading" role.

The test can be compared to a story; the tester is both narrator and actor. As the situation is a stereotype, the responses required are highly predictable. The teacher sets the tasks and has several pupils give the answers. He can note down the results in some way or other. The students can also practice among themselves and write down which tasks they were not able to perform. When a number of tests have been dealt with, it may turn out, for example, that most of the students cannot satisfactorily exchange greetings. The teacher then consults the "social" index and finds out in which of the other tests "greeting" is also included.

If in the "camping" test the students have shown they are good at asking for information, the teacher can check the index for other tests in which "asking for information" also occurs. He can then check whether the students are also able to ask for information in other situations and with reference to other specific notions.

After sufficient practice, the teacher can go through the whole "camping" test with a small separate group of students and give them marks according to two aspects: Was the student successful in getting the message across? Is what has been said formally correct? At this stage the teacher's purpose is to trace shortcomings, so a "soft" form of testing is sufficient, aimed at acquiring feedback for both teacher and student. Our team is presently working on the development of an elaborate rating scale. This presents us with enormous problems, such as determining what specific criteria have to be taken into account in judging communicative ability.

Expectations for the Future

The development at CITO of speaking proficiency tests is still in its initial phase, but already teachers have shown interest in this kind of testing. A few tests have been pretested on a limited scale and experience confirms that the tests meet a long-felt need.

In a year's time the set of tests will be published and, after they have been in use for one or two years, CITO will develop a final test that will be representative of the language behavior as it is described in the area of required language behavior. We hope that the use of these tests will contribute to new developments in foreign language teaching. Speaking the language in class will not be artificial but more practical and true to life. The students will then be motivated, because they will find that they can really "do something" with the language. What they have learned at school will enable them to make contact with foreigners and to communicate with them, both in their own country and while traveling abroad. They will be able in everyday life to say relevant things instead of, for example, giving hardly intelligible expositions on the works of Sartre.

References

Coste, D.; Courtillon, G.; Ferenczi, V.; Martins-Baltar, M.; and Papo, E. *Systèmes d'apprentissage des langues vivantes par les adultes: Un niveau-seuil*. Strasbourg: Conseil de la coopération culturelle du Conseil de l'Europe, 1976.

Council for Cultural Cooperation of the Council of Europe. *Systems development in adult language learning*. A European unit/credit system for modern language learning by adults. Strasbourg, 1973.

v. Ek, J. A. *The Threshold Level*. A European unit/credit system for modern language learning by adults. Strasbourg: Council for Cultural Cooperation of the Council of Europe, 1975.

SCOPE AND LIMITATIONS OF INTERVIEW-BASED LANGUAGE TESTING: ARE WE ASKING TOO MUCH OF THE INTERVIEW?

Robert Lado

Georgetown University

SCOPE AND LIMITATIONS OF INTERVIEW-BASED LANGUAGE TESTING: ARE WE ASKING TOO MUCH OF THE INTERVIEW?

Robert Lado

Introduction

The Physician's Interview and Examination

What happens when one goes to the doctor for a serious examination? The doctor begins by interviewing the patient: "How do you feel? What seems to be the problem? How long has that been bothering you? Have you had those symptoms before? When does it hurt? How is your appetite? Are you able to sleep at night? What is your normal weight? Have you been losing weight lately?" And so forth.

Your attitude is one of cooperation with the physician; that is, you do not try to mislead the doctor or hide your symptoms. Yet, as a rule, the doctor does not make a serious, final diagnosis directly from the interview and first-hand observation of your appearance and behavior. Questions are raised in the doctor's mind. Mental notes are made as the interview proceeds. Hypotheses develop and are often discarded to make way for other possibilities.

Depending on the observations made during the interview, the doctor proceeds with a number of specific tests. The doctor or a trained nurse takes your exact weight instead of accepting your report or making an estimate from your height and the look of your waistline. Stunt men at carnivals can make remarkably accurate guesses of your weight by simply looking at you, and they bet they can guess within five pounds of it or you win a prize. Yet your physician asks you to step on the scale and measures your weight to within a pound or less. The carnival estimator bets he can come within a ten-pound range, and he does not always win. The physician would not even consider recording a sharp-eye estimate.

In addition, the doctor may listen to your heart, check your pulse, or listen through a stethoscope as he taps your chest. He or she does not just hold a tight grip on your arm to estimate your blood pressure as circulation begins to pulse through. A sphygmomanometer measures that pressure so a reading can be made from the height of a mercury column against a scale or from a needle pointing to a circular scale. And notice that to take your pulse rate the physician or the nurse looks at a watch as a count of the pulsations is made. It is easy to train yourself to count seconds quite accurately, yet physicians prefer to look at watches.

The doctor may take a chest X-ray and examine it, make an electrocardiogram, tap your knee for reflexes, and look at your throat, ears, and nose. If there is a hearing problem, the doctor does not just whisper to see if you hear; he or she asks for an audiology test, which measures responses at different sound frequencies.

The physician may take one or more blood samples, or collect a urine specimen, which will be sent to a laboratory and tested for sugar, infection, albumen, or whatever.

Only after the doctor has collected the results of the various specific tests and interpreted them together with the interview does he or she attempt to reach a final diagnosis and prescribe treatment. If the results are inconclusive or contradictory, additional tests are ordered. Where would modern medicine be if doctors depended exclusively on the interview and direct observation of patients?

The Oral Interview Test (OIT)

In the OIT, a trained linguist (the counterpart of the physician) elicits samples of speech by asking questions, suggesting topics, and probing into the usage of the examinee. The examinee, often in an antagonist role, tries to exhibit the best usage and avoid pitfalls that might lower the rating. The trained examiner keeps on probing until satisfied that the true level of performance has been established or until time becomes a problem.

Unlike the medical examination, the OIT does not lead to additional tests to obtain a more complete picture of competence in specific areas. Instead, the examiner searches for questions and topics that might elicit desired responses and exhibit weaknesses and competencies.

It may be argued that linguistic competence is less complex than the functioning of the human body, yet linguistic competence is one of the most complex achievements of a human being. In research on linguistic geography, it takes interviewers many hours of exploration with the aid of questionnaires to report the speech characteristics of a single informant.

By contrast, in an OIT, which lasts from five to thirty minutes, the examiner immediately reaches the diagnosis or rating that says what the examinee can and cannot do in and with the language, or, to use the Civil Service ratings, that the examinee is native-bilingual, full professional, minimum professional, limited working, or elementary in speaking.

From this observation of the examinee in conversation, the examiner decides finally and irrevocably if the examinee can perform full professional functions through the language. And, because of the fact that there is a face-to-face conversation, the examination is considered a valid replication of professional function, which it is not.

When the physician suspects there might be a problem related to the weight of a patient, he or she reaches for the exact weight measurement and does not trust an approximate estimate. The oral interview examiner, however, trusts the approximate estimate. When the physician suspects a hearing problem, he or she does not stop with direct observation, but studies the audiogram showing thresholds at various frequencies on the sound spectrum. And the physician puts more trust in the audiogram, which separates the elements of sound into frequencies, than in the integrative informal test of speaking to determine if the patient hears normally or not. Yet, in the OIT, the examiner does not use any specific measures beyond direct observation of the behavior of the examinee because it is supposedly more valid to do so than to seek more precise information by means of additional tests of various elements. Where can language examinations go if we insist on exclusive reliance on direct impression examinations for our final diagnoses?

Evaluation

So far we have argued only by analogy, and analogy does not prove anything. But we would have to be blindfolded not to recognize that the analogy raises some interesting questions about the possible limitations of the technique. It seems to me that we are justified in assessing in a more formal way the fundamental strengths and weaknesses of the OIT. In testing terms, this means inquiring formally into the validity, reliability, scorability, representativeness, and practicality of the test, and determining what it does and does not do well and how it can be modified or combined with other techniques to produce better results.

Validity

Validity is the most important single criterion to evaluate a test. It is critical because without validity all other criteria, including reliability, are worthless. Validity simply asks whether and to what extent a test measures what it claims to measure. There is no absolute and final answer to the question of validity, since a test only samples what it purports to test. Instead, we search for evidence that supports or weakens its claim, and then, on the basis of all the evidence, we make a judgment.

There are many ways we can seek evidence to answer the validity question. Some of the most convincing evidence comes from (1) face validity, (2) content-of-sample validity, (3) native speaker performance, and (4) empirical or statistical validity.

FACE VALIDITY. The greatest strength of the OIT is its surface or face validity, i.e., the appearance on simple inspection that it tests speaking, which is what it claims to test. The OIT has all the appearance of testing speaking ability: it is actually a speaking performance on the part of the examinee and a speaking performance is not a substitute for speaking but speaking itself.

If we were to rely on face validity alone, we would give the OIT the highest validity rating as a speaking test. Such a rating would be amply justified if speaking a language were as simple as riding a bicycle or driving an automobile. By analogy, the OIT would be equivalent to the road test of a driver's examination.

But mastering a language is more complex than driving a car, and on the basis of the questions raised by our analogy with the physician's examination, we should go beyond face validity into a deeper evaluation of the OIT. Even in a driver's examination, it is common practice to take a written test prior to the road test. And the road test itself is not merely driving around the nearest block but a series of tasks that probe the competence of the driver in various maneuvers.

With regard to the OIT, we notice immediately that it is a restricted sample of speaking that, as such, may or may not give a fully accurate picture of linguistic or communicative competence. This leads us to content-of-sample validity.

CONTENT-OF-SAMPLE. Content in a language test refers to the language and the situations tested. We know that language is a system of rules, patterns, and lexical items and their meanings used by a speech community to communicate and interact in carrying on the multiple functions typical of life in that community. We should, therefore, inquire into the content of the OIT with regard to grammatical system, vocabulary, pronunciation, situations, and fluency.

Grammatical System. In the OIT the examinee may not have sufficient opportunity to ask questions, for example, or to use requests, invitations, or exclamations, or use various types of complex sentences or passive or reflexive constructions. The experienced examiner guards against such lacunae but may not be able to elicit utterances containing important elements of competence such as the different types of questions, including those of the yes/no, information, subject, verb phrase, predicate, and echo types, among others.

We all agree that the total language system cannot be tested in one interview and that we must, therefore, be satisfied with a sample. But how is that sample to be chosen? By subjective impressions? By error counts? By linguistic analysis? Without precise criteria concerning the sample, there is bound to be variation among interviewers and from one interview to another with regard to the elements elicited. An informal general list such as examiners often have in mind allows too much variation.

In a recorded OIT of Spanish, which lasted twenty minutes and yielded an S rating of 4, the examiners asked fifty-five questions and the examinee none (DeCesaris, 1977). The examiners made a clear effort to elicit the subjunctive and conditional forms, but they overlooked the area of interrogatives completely.

Years ago, I was called as a consultant to evaluate an OIT under development for the Air Force to test illiterate Puerto Rican recruits in spoken English. It was a carefully structured interview that sought to test competence in a number of areas. On examining it, I discovered that it did not provide for questions to be put by the examinee.

Vocabulary. We know that even full bilinguals do not have completely parallel competence in all lexical areas of the two languages. I, for example, feel less competent to discuss psychology in Spanish than in English, because practically all my study of psychology was in English, but I feel more competent to discuss literature in Spanish. Should the topic be soccer, I would again do better in Spanish; if it were current movies, I would do badly in both. Yet, on the basis of a conversation on some informally chosen topic, the OIT may report a rating of S-4, full professional proficiency, which is described as "able to use the language fluently and accurately on all levels normally pertinent to professional needs," without necessarily sampling the lexical areas in which full professional competence has been achieved.

Pronunciation. The OIT provides a highly valid sample of an examinee's competence in pronunciation, with respect to both face validity and content-of-sample validity. Practically all the phonemes and phoneme sequences of the language and most of the intonation and rhythm patterns will be exhibited. There are problems with regard to scoring, but not with validity.

Situational Content. One of the strengths of the OIT is that it represents performance in a communicative situation. This is more valid than reciting memorized texts as a measure of speaking, and it is more valid than a repetition test described by Politzer et al. (1974). It is more valid than the noise test, which is essentially a dictation with noise interference, as reported in Spolsky et al. (1968) and Gaies et al. (1977).

By attempting to introduce different questions and tasks, the examiner tries to improve the situational content. In this sense the OIT can be more effective than a picture stimulus test if the examiners are experienced. Nevertheless, the OIT is not fully representative for two reasons. (1) The OIT is a test of conversational competence rather than of extended formal speaking. It does not sample the ability of a professor to deliver a lecture to a class, or of an ambassador to give a public lecture, as ambassadors are often invited to do. (2) It does not sample sociolinguistic variations, which are sometimes critical in effective communication. Notice, for example, variations required in addressing men and women, older and younger persons, individuals of high status, and in-house employees of different sociolinguistic status. Of course, these differences could be deliberately sought out in the interview and become part of it. The question would be then whether the OIT were too long. Would its spontaneity be hampered? Could these variations be tested by other means?

Fluency. Fluency is sampled quite adequately in the OIT. As with pronunciation, any problems with regard to fluency will be in scoring rather than in validity. Are all examiners rating the same thing when they rate fluency? Should it be more explicitly defined?

NATIVE SPEAKER PERFORMANCE. The OIT seems strong with regard to native speaker performance. All examinees would presumably perform at a rating level of 5 if tested in their native language. Yet there is one area that leaves some doubt in my mind. It is the matter of poise, personality, and presence. Would all examinees give a typical performance each time if tested in their native language? We are intuitively aware that we do not always perform at our best under all circumstances. Is there any substance to this impression?

Differences in performance among educated adults may not turn out to be of major importance, but differences among children are substantial, as reported by sociolinguistic studies of ghetto children. I recently made a sound movie of a two-year-old Spanish-speaking child learning to read Spanish. The parents had reported that he was able to read three books of an experimental series. Yet, when we attempted to film his performance, he did not read a single word, even though the filming was at home with his parents. The OIT is not a test for two-year-olds, of course, but it would be interesting to test some adult examinees in their native language to see what performance they actually display.

EMPIRICAL VALIDITY. A standard empirical validation of a test is its correlation with a valid criterion. The valid criterion could be the scores on a speaking test whose validity has been previously established. With the OIT we cannot use this approach because we simply do not have a fully validated and established speaking test.

To obtain a more valid criterion, we will have to turn to (1) a more extended version of the OIT with adequate sampling of situations and language, (2) an increase in the number of graders or an increase in their competence, or (3) a combination of the above. If it turns out that the OIT correlates highly with the longer and better-structured version scored by a group of qualified examiners, we would be justified in considering the OIT validated.

I have not seen such a validation attempt. Instead, I have seen a proposal that a shorter version be correlated with the full OIT to validate the shorter version. Obviously, if the shorter version correlated highly with the normal length OIT, we would gain by the practical advantage of its shortness.

However, since we are still exploring possible limitations of the OIT, its validation with a longer, structured OIT scored by more than two judges would seem to be of greater interest. Another possibility is the use of in-depth interviews supplemented by additional tests.

Reliability

Reliability has to do with the stability of obtained scores. If scores fluctuate excessively for the same students on repeated administrations, the test is unreliable. The extent to which scores are reliable is expressed as a correlation between two sets of scores made by the same students on the same test. In reliability, then, the test is correlated not with a separate criterion, as in empirical validity, but with itself.

The fewer the possible grades on the scale of a test, the easier it is to attain high reliability. The extreme case is a pass-fail test with a single cutoff point between passing and failing. Most students will be either far above or far below the cutoff point and thus assure high reliability since only those that are close to the cutoff point are likely to fluctuate.

The OIT rating scale is based on nine effective slots, 0+, 1 and 1+, 2 and 2+, 3 and 3+, and 4 and 4+. It is not difficult to attain high reliability with such a scale. If scores were distributed over fifty or a hundred points on the scale, we would expect the reliability of the OIT to be lower.

The nine-point scale is apparently satisfactory for present government users of the test. For academic purposes, however, it is too coarse and tends to bunch up scores around the 1 and 1+ ratings, masking progress within and between them. The nine-point scale is a weakness also for control-type research because it tends to flatten out significant differences in achievement in the range where most scores fall.

Wilds (1975), while staunchly affirming, "The fact of the matter is that this system works," admits that

> Even in languages in which tests are conducted frequently as French and Spanish, where there is no doubt that standards are internalized and elicitation techniques are mastered, it is possible for criteria to be tightened or relaxed unwittingly over a period of several years so that ratings in the two languages are not equivalent or that current ratings are discrepant from those of earlier years.

and

> It is, however, very much an in-house system which depends heavily on having all interviewers under one roof, able to consult with each other and share training advances in techniques or solutions to problems of testing as they are developed and subject to periodic monitoring. It is most apt to break down as a system when examiners are isolated by spending long periods away from home base (say a two-year overseas assignment), by testing in a language no one else knows, or by testing so infrequently or so independently that they evolve their own system. (p.35)

The fact that two examiners are required to rate the OIT indicates lack of confidence in the rating by one examiner. This compares unfavorably with standard practice in testing, which as a rule relies on one scorer. Because of weaknesses in reliability, the practice of using two examiners should be maintained if practical from the point of view of trained personnel and cost. Dyson (1972) found that a shorter examination with team marking was better than a longer test with a single marker.

Scorability

The subjective nature of the OIT scoring is one of its weaknesses in its present form and use. According to Clark (1975), it takes four full days to train an examiner. And Wilds (1975) indicates, as quoted above, that examiners who are out in the field for two years must be retrained. The CIA has its two examiners rate the interview separately, and averages the ratings on a scale. The FSI has the interviewers discuss their differences to arrive at an agreement. These are indications that scoring the OIT is difficult and subjective to a significant degree. Improvement in this area is obviously desirable.

A standard way to improve objectivity in scoring is to identify the measurable parameters of competence. The rating scales for accent, grammar, vocabulary, fluency, and comprehension reported by Wilds (1975) represent an effort in this direction. One may be puzzled, however, by the weights of the different components: three points to grammar, two to vocabulary, one to fluency, two to comprehension, and zero to accent.

This cannot mean that pronunciation is not an important factor in speaking. Pronunciation contributes to intelligibility even though redundancy resolves many inaccuracies in pronunciation . Furthermore, sociolinguistic studies show that foreign language accentedness and social dialect markedness are perceived and judged by native speakers very quickly. A speaking test must, therefore, be considered incomplete until pronunciation is taken into account, either on a complex scale showing foreign and social dialect dimensions or on an inventory of pronunciation features or phonemes and sequences. And if this makes the OIT too difficult to score by available examiners, it should be supplemented with a pronunciation test of some kind to give us a better picture of speaking skill.

Practicality

Practicality must be considered in conjunction with the particular uses intended for the OIT. The FSI, CIA, Peace Corps, and other agencies and organizations that have the trained personnel on hand and can keep careful control of ratings find the OIT practical. The estimated cost of $35 per examination (Jones, 1975, p.9) and the fifteen interviews that can be administered by a team of two examiners in a working day (Clark, 1975, p.20) are also acceptable to those users. A twenty-minute interview by two trained examiners limits the use of the OIT in university and high school settings for practical reasons. It would take a team of examiners a full working week and two additional days to test 100 students, a not uncommon task in those settings.

If the OIT were shortened to, say, five minutes, its practicality would be significantly enhanced. If, in addition, a single examiner were used, subject to checking by a second examiner when challenged, a further improvement in practicality would be effected.

The OIT as a Listening Comprehension Test

The OIT shows obvious weaknesses as a listening comprehension instrument. In the interview that I analyzed from a recording, the examiners asked fifty-five questions and the examinee required clarification only once. In speaking, however, the examinee did not ask any questions. The speaking sample was exclusively expository and narrative. In listening comprehension it was all questions and no narration or exposition. This represents a weakness in content-of-sample validity. Furthermore, it is doubtful that any careful check could have been kept on comprehension, since attention was on speaking.

Kaufman (1969) compared the S-ratings of forty-four Peace Corps volunteers on the OIT with their listening comprehension scores on the Pictorial Auditory Comprehension Test (PACT) developed for the Peace Corps by John B. Carroll. PACT is a seventy-five item multiple-choice test that uses four pictures as alternatives for each item. The tests were administered after a nine-week intensive course in Spanish conducted in Puerto Rico. The interviews were administered by Kaufman shortly after he was recertified by the Foreign Service Institute to administer the OIT in Spanish to Peace Corps volunteers. Kaufman was assisted throughout the oral testing by a Puerto Rican and a Colombian, who had not been involved in the training of these volunteers.

The S-ratings on the OIT and the listening comprehension (LC) scores on PACT are presented in Table 1. The correlation between the two sets of scores, using the Pearson product-moment linear correlation formula, was .83. This is fairly high and could be used to compare performances by groups of similar students. Looking into a comparison of performance by individuals, however, a different picture emerges.

Dividing the PACT scale into nine intervals to parallel the nine OIT ratings, and equating the two scales at their modes, (the slots with the largest number of scores in each scale), we note that 68 percent of the students who rated within the five levels 0, 1, 2, 3, and 4 (without separating the 0+, 1+, etc.) also rated within the corresponding double intervals on the PACT scores, while 32 percent were either above or below. Using the full nine-point scale on both the OIT and PACT, 36 percent of the students remained in the same slot and 64 percent were either above or below.

TABLE 1

Spanish OIT S-Ratings & PACT LC Scores of 44 Peace Corps Volunteers

	PACT LC Scores	4+	4	3+	3	2+	2	1+	1	0
	71	4+				2+				
(4+)*	70		4							
	69			3+						
	68									
(4)	67						2			
	66									
	65				3					
	64					2+				
	63									
(3+)	62									
	61						2			
	60									
	59									
	58									
(3)	57						2			
	56									
	55						2			
	54									
	53								1	
(2+)	52							1+		
	51									
	50									
	49									
	48								1 1	
(2)	47							1+		
	46						2		1	
	45								1 1	
	44								1 1 1 1	
	43								1 1	
(1+)	42								1	
	41							1+	1	
	40								1 1 1	
	39								1	
	38								1	
(1)	37									0+
	36								1 1 1	
	35								1 1 1	
	34								1	
	33									
(0+)	32								1 1	
	31									
	30									
	...									
	20								1	
	OIT S-RATINGS	4+	4	3+	3	2+	2	1+	1	0

*Indicates what the LC rating would have been if measured by PACT.

In other words, if we use the OIT speaking ratings to predict PACT listening comprehension performance using a nine-slot rating scale, we are off by at least one level in approximately two-thirds of the cases, indicating that the OIT S-ratings are not satisfactory measures of listening comprehension. The reverse would also be true; that is, if we use PACT listening comprehension scores to predict speaking performance in terms of OIT ratings, we are off by at least one level in approximately two-thirds of the cases, indicating that PACT listening comprehension scores are not valid measures of speaking performance. This is further confirmed by looking at some specific cases. We notice, for example, that one student rated 2+ by the OIT would be rated 4+ by PACT. Another student, with OIT 2, would rate PACT 4. And a third student, with OIT 1, would rate PACT 2+.

Consequently, since a listening comprehension test can be administered with ease to individuals as well as groups by examiners with standard training, and since results are scored objectively and quickly, separate listening comprehension tests are to be preferred in all cases in which examinees are willing to submit to them.

What the OIT Does and Does Not Do Well and What to Do about It

Selecting and condensing some of the above considerations, it is not unreasonable to state the following conclusions and recommendations.

1. The OIT is the best available test to obtain a valid speaking sample. It should, therefore, be retained when the necessary requirements with regard to personnel training and availability and budget provisions are present.

2. The representativeness of the speaking sample is less satisfactory than that of professionally prepared tests of listening comprehension, reading, and writing. Therefore, the OIT should be further structured to ensure better sampling of linguistic, situational, and sociolinguistic components, or it should be supplemented by other tests that are more effective in those areas. The OIT could then be shortened to a more practical and uniform length.

3. Scoring of the OIT is unusually difficult and must be presumed uneven under ordinary testing conditions. This problem can be minimized by not relying exclusively on the OIT but supplementing it instead with other objective tests.

4. The OIT is not a good test of listening comprehension by psychometric standards. It should, therefore, not be used as a measure of that skill. Listening comprehension tests are far superior and can be administered individually as well as in groups at a fraction of the cost of the OIT and with lower demands on personnel training.

5. The OIT is not a practical test of competence on internalization of grammar, vocabulary, and pronunciation, because of sampling and scoring problems. Therefore, it should be supplemented whenever possible with tests of those components when they are deemed necessary.

6. The OIT is not a test of reading or writing and should not be used as a measure of those skills. This is stated to counter any claim that language competence is general in nature and need not be tested in its different manifestations.

7. Since the OIT is difficult to administer and score, and because it requires highly trained personnel not always available, it should be restricted to VIPs who might not be willing to submit to other types of tests. For wider use, a short version of the OIT with more limited goals, supplemented by additional tests, is recommended.

Conclusion

To the query whether we are asking too much of the OIT in its present form, the answer is yes. Therefore, we should either ask less of the interview and supplement it with tests that are better adapted to some of the components, or, rejecting that, we should extend the interview and structure it so it will provide a better sample of linguistic, situational, and sociolinguistic competence.

More specifically, in this observer's opinion, we should keep the OIT since it is a valid test of speaking and supports teaching and evaluation of speaking, but we should make it shorter, more uniform in length, and supplement it with tests of listening comprehension, reading, grammar, vocabulary, pronunciation, and writing for a more complete picture of competence. We should also increase the number of subcategories under each rating so as to reflect more adequately the vast achievement that mastery of a second language represents.

References

Beardsmore, H. Baetens. "Testing Oral Fluency." IRAL 12 (1974): 317-25.

Clark, John L. D. "Theoretical and Technical Considerations in Oral Proficiency Testing." In Testing Language Proficiency, edited by Randall L. Jones and Bernard Spolsky, pp. 10-24. Arlington, Va.: Center for Applied Linguistics, 1975.

Coward, D. A. "Confessions of an Oral Examiner." Modern Languages 68 (1977): 35-38.

Davison, J. M., and Geake, P. M. "An Assessment of Oral Testing Methods in Modern Languages." Modern Languages 51 (1970): 116-23.

DeCesaris, Janet. "The FSI Interview." Unpublished term paper with cassette recording, Georgetown University, 1977.

Dyson, A. P. "Oral Examining in French." Modern Language Journal 53 (June 1972): 54-55.

Gaies, S. J.; Gradman, H. L.; and Spolsky, B. "Toward the Measurement of Functional Proficiency: Contextualization of the Noise Test." TESOL Quarterly 11 (1977): 51-57.

Johansson, S. "An Evaluation of the Noise Test--A Method for Testing Overall Second Language Proficiency by Perception Under Masking Noise." IRAL 11 (1973): 107-33.

Jones, Randall L., and Spolsky, Bernard, eds. Testing Language Proficiency. Arlington, Va.: Center for Applied Linguistics, 1975.

Kaufman, David. "Comparison of Speaking Proficiency with Auditory Comprehension--An Experiment." Unpublished term paper, Georgetown University, 1969.

Politzer, Robert; Hoover, Mary Rhodes; and Brown, Dwight. "Test of Proficiency in Black Standard and Nonstandard Speech." TESOL Quarterly 8 (1974): 27-35.

Rey, Alberto. "A Study of the Attitudinal Effect of a Spanish Accent on Blacks and Whites in South Florida." Unpublished doctoral dissertation, Georgetown University School of Languages and Linguistics, 1974.

Shuy, Roger W. "Sociolinguistics." In Linguistic Theory: What Can It Say about Reading?, edited by Roger Shuy, pp. 80-94. Newark, Del.: International Reading Association, 1977.

Spolsky, Bernard; Sigurd, Bengt; Sako, Masahito; Walker, Edward; and Arterburn, Catherine. "Preliminary Studies in the Development of Techniques for Testing Overall Second Language Proficiency." Language Learning 18 (August 1968): 79-101.

Wilds, Claudia P. "The Oral Interview Test." In Testing Language Proficiency, edited by Randall L. Jones and Bernard Spolsky, pp. 29-38. Arlington, Va.: Center for Applied Linguistics, 1975.

MEASURING FOREIGN LANGUAGE SPEAKING PROFICIENCY:
A STUDY OF AGREEMENT AMONG RATERS

Marianne L. Adams

Foreign Service Institute

MEASURING FOREIGN LANGUAGE SPEAKING PROFICIENCY: A STUDY OF AGREEMENT AMONG RATERS[1]

Marianne L. Adams

Background

Proficiency in speaking a foreign language is more often inferred than directly measured. Perhaps this is because of the difficulty of scoring speaking examinations objectively. Yet, in an organization whose purpose it is to communicate with foreign nationals, foreign language proficiency must be measured, because inferring a person's speaking proficiency from the person's ability to read, write, or listen may not be valid. Although the assessment of speaking proficiency is difficult, the responsibility is unavoidable.

The School of Language Studies at the Foreign Service Institute (FSI) trains and tests government employees for overseas service. The purpose of the testing program is to provide information about the professional usefulness of a given person's knowledge of a language. "How much of the business of the United States government in country X would the employee be competent to do in language X?" is the question FSI attempts to answer. One key feature of the testing program is that employees are assigned to "proficiency levels" based on their oral test performance. Employee proficiency level assignments are based on the match between an employee's oral test performance and prespecified levels of performance required for each proficiency level. Therefore, the Foreign Service Institute language proficiency test is referred to as a "criterion-referenced test."

The speaking portion of the FSI language proficiency test consists of an oral interview structured with reference to the proficiency levels. The candidate is always asked to converse with a native speaker of the target language on topics as complex as he or she can manage. Three people take part in the test: the candidate, an interviewer, and an examiner. The last is in charge of the test and, while mostly the examiner listens, occasionally he or she directs the conversation.

Criterion-referenced tests are often contrasted with the better known norm-referenced tests. A norm-referenced test is constructed and used principally to facilitate making comparisons among individuals on the ability measured by the test. Clearly, a norm-referenced test would not meet FSI's needs. Because the purpose of a criterion-referenced test--to provide a clear description of what a candidate can do--is fundamentally different from that of a norm-referenced test, it is not surprising that methods for test development and evaluation differ considerably for the two types of tests (Hambleton and Novick, 1973; Millman, 1974; Swaminathan, Hambleton, and Algina, 1974).

[1] The author would like to acknowledge the helpful comments and constructive criticisms of Ronald K. Hambleton of the University of Massachusetts, Amherst.

The test is widely used and enjoys a good reputation. It has been adopted by organizations faced with the need for speakers of foreign languages, e.g., the Peace Corps and some businesses. The test has both content validity and face validity and a clientele that has substantial confidence in the reliability of the proficiency ratings. Nevertheless, there is an ongoing need for technical analyses of the test and its characteristics.

The study reported here was designed to address the problem of agreement among different raters of proficiency level assignments to the same set of candidates. Specifically, the study was designed to address the following questions:

1. Could the selection of a rater influence proficiency level assignments (and if so, by how much)?

2. What would be the nature of disagreements in ratings? (For example, do disagreements in ratings between two examiners follow a random pattern?) Also, since some disagreements are more serious than others (mastery-nonmastery determination), what percentage of the time do raters agree in their mastery or nonmastery determination of candidates?

3. How do the results from questions 1 and 2 above compare for tests in three languages: French, German, and Spanish?

These questions refer, of course, to only one aspect of the test: the individual rater. In the actual work situation, however, no rater judges a test alone. Raters always work in pairs. The pairs of raters also work under well-defined testing procedures and criteria of the test.

The results are underestimates of true reliabilities because many of the inconsistencies are removed by consultation. In this study we let inconsistencies stand.

Definitions

At this point it will be useful to define several terms:

1. Oral Interview--A test of speaking proficiency in a foreign language.

2. Foreign Language--There were three foreign languages of interest in this study: French, German, and Spanish.

3. Proficiency Scale--The scale consists of eleven points: 0, 0+, 1, 1+,..., 4, 4+, 5. The labels attached to six of these points are as follows:

0 - No Proficiency
1 - Elementary Proficiency
2 - Limited Working Proficiency
3 - Minimum Professional Proficiency
4 - Full Professional Proficiency
5 - Native or Bilingual Proficiency

If proficiency substantially exceeds the minimum requirements for the level involved but fails short of performance required at the next higher level, a "plus" is attached to a candidate's proficiency level.

4. Mastery Status--Besides the eleven proficiency levels, an important distinction is made between persons scoring 3 and above and those scoring below 3. For purposes of this paper, I call persons receiving scores 3 and above "masters" because there are certain professional rewards in the U.S. Foreign Service for proficiency at the 3 level and above. I call others "nonmasters." (Disagreements between examiners that affect the "mastery status" of persons are far more serious than disagreements that do not.)

5. Testing Team--Consists of two raters, one known as examiner and one known as interviewer.

The interviewer is usually a native speaker of the language being tested and has received training in conducting FSI test interviews. The examiner is linguistically oriented in one or more foreign languages, including the one being tested. He or she is in charge of the administration of the test. This responsibility includes instructing the interviewer on the line of questioning, setting hypothetical role-playing situations, supplying stimuli for conversation, and discussing the test results with the candidate.

The examiner and the interviewer have equal voices in rating a test. They vote on the results of a test. If their opinions differ by half a point, the lower grade is awarded. If their opinions differ by a full point, they submit their test, tape, and notes to arbitration by the head of the testing unit.

Interviewers did not always have an equal voice in the grading decision; rating was added to their duties just prior to this study. The results of this study for them must be considered in light of the novelty of the task.

Procedure

Examiners and interviewers in French, German, and Spanish listened individually to tapes of fifty tests (oral interviews) and rated them independently.[2] The complete list of participants is included in Appendix A. In total, we had six in French, four in German, and eleven in Spanish. Four to six tapes at each of the eleven proficiency levels were selected for use in the study, with the exception of level 5, where only two or three examples per language were selected. By allowing the number of tapes to vary, we prevented the participants from determining a proficiency level based on an expected number of cases.

Some tapes had to be withdrawn from the study for lack of acoustic fidelity. The final count of tapes used in the study was as follows: French--fifty, German--forty-six, and Spanish--forty-eight.

Several raters did not judge every test. Others gave more than one rating to some tapes. Fortunately, the numbers of times these events occurred was very small. Rather than disqualify the raters, the investigator supplied the average of grades given by other raters to fill the gaps.

In total, five raters did not rate a complete set of tapes. The situation was as follows:

Rater	Number of Tapes Rated
French, D	49
German, C	45
German, D	45
Spanish, I	47
Spanish, J	45

The ratings were completed in two time periods:

1974 - Examiners	1977 - Interviewers
French - raters A and B	French - raters C, D, E, and F
German - raters A and B	German - raters C and D
Spanish - raters A, B, C, D, and E	Spanish - raters F, G, H, I, J, and K

Results and Discussion

In our first analysis, we correlated the ratings of each pair of examiners across the approximately fifty tapes. The correlations between the pairs of examiners for the French, German, and Spanish raters are

[2] Twenty-five of the interviews were recorded at FSI and twenty-five at the CIA Language School as part of a joint project between the two schools.

reported in Tables 1, 2, and 3. (The ratings data from which the correlations were computed are reported in Appendix B.) It is clear from the tables that there was a nigh level of agreement among the raters. Correlations between their ratings in all cases exceeded .82, with the average correlation .91.

The correlations reported in Tables 1, 2, and 3 are even more impressive when one considers that the tapes presented each rater with a possible range of eleven choices of ratings for each test (the more possible choices of ratings, the more room for disagreement among the raters). The high correlation coefficients show that there was substantial agreement among the raters as to the criteria.

Correlation tables are an interesting by-product but not the central thrust of this study. For our purposes, we were more interested in the kinds and degrees of disagreements--whether raters tended to assign approximately the same ratings, or whether some were overly generous and others overly strict.

TABLE 1

Pair-Wise Correlations* of French Testers' Ratings of the Tapes

Rater	Rater				
	B	C	D	E	F
A	.95	.92	.92	.93	.93
B		.92	.92	.90	.92
C			.94	.89	.93
D				.92	.95
E					.96

TABLE 2

Pair-Wise Correlations* of German Testers' Ratings of the Tapes

Rater	Rater		
	B	C	D
A	.89	.93	.93
B		.87	.88
C			.98

*Pearson product-moment correlation coefficients.

TABLE 3

Pair-Wise Correlations* of
Spanish Testers' Ratings of the Tapes

Rater	B	C	D	E	F	G	H	I	J	K
A	.95	.95	.96	.92	.94	.88	.91	.89	.94	.89
B		.96	.96	.95	.92	.92	.94	.89	.94	.91
C			.96	.91	.91	.90	.89	.92	.94	.91
D				.93	.93	.91	.91	.94	.95	.91
E					.90	.87	.95	.85	.91	.90
F						.87	.88	.87	.91	.94
G							.84	.88	.92	.82
H								.83	.92	.91
I									.90	.87
J										.90

*Pearson product-moment correlation coefficients.

Tables 4, 5, and 6, corresponding to the French, German, and Spanish raters' data, respectively, summarize several pieces of pertinent data for the purpose of this study. For the French raters, the average percentage of ratings in agreement or tolerable disagreement was 92 percent. The average percentage of times raters agreed on a candidate's mastery status was 92 percent. Average percentage of agreement for the Spanish raters was 87 percent and agreement on mastery status was 94 percent.

Table 7 shows that the errors in proficiency level determination that do occur were, for the most part, not patterned. Only one rater was consistently more generous, and one was consistently more severe.

What does it all mean? We would obviously like to have perfect agreement, but every improvement has its price.

There are several known ways to increase reliability: reduce the number of points in the scale, reduce the number of raters, lengthen testing time. If we reduce the scale, we sacrifice information. If we reduce the number of raters, we might overburden those who do test and thus introduce a further error component. If we increase testing time, we increase the cost.

TABLE 4

An Analysis of Proficiency Level Assignments for Each Pair of French Raters

Rater Pair	Number of Tapes	Percentage of Ratings in:			Identical Mastery Status[d]
		Perfect Agreement[a]	Tolerable Disagreement[b]	Total Agreement[c]	
*A,B	50	78	16	94	94
A,C	50	76	20	96	96
A,D	49	52	35	87	88
A,E	50	74	16	90	92
A,F	50	74	14	88	88
B,C	50	64	22	86	94
B,D	49	60	24	84	90
B,E	50	64	24	88	86
B,F	50	70	20	90	90
**C,D	49	51	37	88	88
**C,E	50	78	18	96	92
**C,F	50	62	28	90	88
**D,E	49	55	37	92	88
**D,F	49	57	35	92	88
**E,F	50	64	24	88	84
Averages					
Examiners		78	16	94	94
Interviewers		61	30	91	88
Actual Teams		67	22	89	91
All Raters Combined		69	23	92	92

[a] Perfect agreement = Percent of identical ratings of a tape by two raters, e.g., rater A's "3" = rater B's "3" or rater A's "3.5" = rater B's "3.5."

[b] Tolerable disagreement = Percent of ratings of a tape by two raters differing by .5 point across whole numbers, e.g., rater A's "3.5" = rater B's "4.0."

[c] Total agreement = "Perfect agreement" plus "tolerable disagreement."

[d] Identical mastery status = Percent of times that two raters agree in their mastery status determination.

*Examiners.
**Interviewers.

TABLE 5

An Analysis of Proficiency Level Assignments for Each Pair of German Raters

Rater Pair	Number of Tapes	Percentage of Ratings in: Perfect Agreement[a]	Tolerable Disagreement[b]	Total Agreement[c]	Identical Mastery Status[d]
*A,B	45	49	36	85	87
A,C	45	62	24	86	87
A,D	45	71	16	87	87
B,C	45	51	22	73	96
B,D	45	56	24	80	84
**C,D	45	93	07	100	100
Averages					
Examiners		49	36	85	87
Interviewers		93	07	100	100
Actual Teams		60	22	82	89
All Raters Combined		67	22	89	92

[a] Perfect agreement = Percent of identical ratings of a tape by two raters, e.g., rater A's "3" = rater B's "3" or rater A's "3.5" = rater B's "3.5."

[b] Tolerable disagreement = Percent of ratings of a tape by two raters differing by .5 point across whole numbers, e.g., rater A's "3.5" = rater B's "4.0."

[c] Total agreement = "Perfect agreement" plus "tolerable disagreement."

[d] Identical mastery status = Percent of times that two raters agree in their mastery status determination.

*Examiners.
**Interviewers.

TABLE 6

An Analysis of Proficiency Level Assignments for Each Pair of Spanish Raters

Rater Pair	Number of Tapes	Percentage of Ratings in: Perfect Agreement[a]	Tolerable Disagreement[b]	Total Agreement[c]	Identical Mastery Status[d]
*A,B	48	73	23	96	98
*A,C	48	82	12	94	98
*A,D	48	73	23	96	94
*A,E	48	73	19	92	94
A,F	48	71	15	86	94
A,G	48	67	21	88	96
A,H	48	75	10	85	92
A,I	47	66	19	85	88
A,J	45	64	27	91	92
A,K	48	58	23	81	88
*B,C	48	79	17	96	98
*B,D	48	65	25	86	94
*B,E	48	71	25	96	98
B,F	48	62	21	83	96
B,G	48	69	15	84	94
B,H	48	65	17	82	98
B,I	47	66	25	91	87
B,J	45	67	24	91	98
B,K	48	42	33	75	90
*C,D	48	73	19	92	96
*C,E	48	75	17	92	94
C,F	48	62	25	87	94
C,G	48	67	21	88	96
C,H	48	71	15	86	96
C,I	47	74	13	87	87
C,J	45	76	16	92	96
C,K	48	52	21	73	90
*D,E	48	71	19	90	94
D,F	48	65	19	84	90
D,G	48	52	33	85	92
D,H	48	77	17	94	96
D,I	47	57	17	74	85
D,J	45	58	31	89	91
D,K	48	58	15	73	88
E,F	48	62	19	81	98
E,G	48	67	19	86	94
E,H	48	67	23	90	94
E,I	47	66	15	81	92
E,J	45	73	20	93	94
E,K	48	50	27	77	90

(Continued on page 140)

TABLE 6 (cont.)

Rater Pair	Number of Tapes	Percentage of Ratings in: Perfect Agreement[a]	Tolerable Disagreement[b]	Total Agreement[c]	Identical Mastery Status[d]
**F,G	48	56	23	79	94
**F,H	48	67	17	84	94
**F,I	47	55	21	76	92
**F,J	45	62	22	84	98
**F,K	48	79	12	91	88
**G,H	48	58	17	75	94
**G,I	47	55	30	85	B8
**G,J	45	69	18	87	96
**G,K	48	48	23	71	90
**H,I	47	55	23	78	85
**H,J	45	62	24	86	96
**H,K	48	52	21	83	92
**I,J	44	57	23	90	86
**I,K	47	40	34	72	83
J,K	45	62	22	84	87
Averages					
Examiners		74	37	92	96
Interviewers		58	45	81	91
Working Pairs		64	42	85	93
All Raters Combined		65	42	87	94

[a] Perfect agreement = Percent of identical ratings of a tape by two raters, e.g., rater A's "3" = rater B's "3" or rater A's "3.5" = rater B's "3.5."

[b] Tolerable disagreement = Percent of ratings of a tape by two raters differing by .5 point across whole numbers, e.g., rater A's "3.5" = rater B's "4.0."

[c] Total agreement = "Perfect agreement" plus "tolerable disagreement."

[d] Identical mastery status = Percent of times that two raters agree in their mastery status determination.

*Examiners.
**Interviewers.

TABLE 7

Direction of Errors among Pairs of Raters

Rater Pair	Number of Tapes	Total Number of Disagreements	Number of Times: First Rater Higher	Second Rater Higher	X^2
			FRENCH RATERS		
*A,B	50	24	12	12	.00
A,C	50	24	12	12	.00
A,D	49	29	9	20	4.18
A,E	50	28	10	18	2.28
A,F	50	23	17	6	5.26†
B,C	50	30	13	17	.52
B,D	49	30	8	22	6.54†
B,E	50	34	9	25	7.52†
B,F	50	28	18	10	2.28
**C,D	49	29	10	19	2.80
**C,E	50	28	10	18	2.28
**C,F	50	28	20	8	3.84
**D,E	49	22	16	6	4.54
**D,F	49	31	26	5	14.22†
**E,F	50	33	26	7	10.92†
			GERMAN RATERS		
*A,B	45	33	19	14	.74
A,C	45	29	12	17	.86
A,D	45	31	20	11	2.30
B,D	45	27	18	9	3.00
B,D	45	29	19	10	2.78
**C,D	45	9	2	7	2.78

*Examiners.
**Interviewers.
† = Significant at p < .05 level.

Table 7 (cont.)

Rater Pair	Number of Tapes	Total Number of Disagreements	Number of Times: First Rater Higher	Second Rater Higher	χ^2
		SPANISH RATERS			
*A,B	48	20	10	10	.00
*A,C	48	26	11	16	.92
*A,D	48	30	13	17	.54
*A,E	48	20	10	10	.00
A,F	48	24	8	16	2.66
A,G	48	27	12	15	.24
A,H	48	21	14	7	1.81
A,I	47	24	14	10	.66
A,J	45	21	8	13	1.20
A,K	48	31	6	25	11.64†
*B,C	48	28	12	16	.56
*B,D	48	24	13	11	.16
*B,E	49	24	11	13	.16
B,F	48	27	12	15	.34
B,G	48	28	12	16	.56
B,H	48	29	15	14	.04
B,I	47	25	14	11	.36
B,J	45	27	9	18	3.00
B,K	48	38	8	30	6.36†
*C,D	48	20	8	12	.8
*C,E	48	21	7	14	2.34
C,F	48	34	12	22	2.94
C,G	48	33	11	22	3.67
C,H	48	25	13	12	.04
C,I	47	19	9	10	.06
C,J	45	21	3	18	10.71†
C,K	48	35	5	30	29.76†
*D,E	48	19	9	10	.06
D,F	48	28	9	19	3.57
D,G	48	37	18	19	.02
D,H	48	21	14	7	2.33
D,I	47	25	13	12	.04
D,J	45	25	10	15	.50
D,K	48	32	7	25	5.06†
E,F	48	23	8	15	2.12
E,G	48	26	11	15	.62
E,H	48	26	10	16	1.38
E,I	47	23	14	9	1.08
E,J	45	17	7	10	.54
E,K	48	30	6	24	10.80†

*Examiners.

**Interviewers.

† = Significant at $p < .05$ level.

Table 7 (cont.)

Raters	Number of Tapes	Total Number of Disagreements	Number of Times: First Rater Higher	Second Rater Higher	X^2
**F,G	48	35	19	16	.26
**F,H	48	33	24	9	6.82†
**F,I	47	32	21	11	3.12
**F,J	45	29	14	15	.04
**F,K	48	26	5	21	9.85†
**G,H	48	34	20	14	1.06
**G,I	47	32	19	13	1.12
**G,J	45	22	10	12	.18
**G,K	48	36	11	25	5.44†
**H,I	47	29	13	16	.30
**H,J	45	26	8	18	3.85†
**H,K	48	34	4	30	19.88†
**I,J	45	23	6	17	5.26†
**I,K	47	36	8	28	5.56†
**J,K	45	26	6	30	7.54†

*Examiners.
**Interviewers.
† = Significant at p<.05 level.

In actual practice the rate of agreement is higher because all possible pairs do not constitute testing teams. The average rate of agreement for actual testing teams in French and Spanish was 89 percent and 85 percent. The average rate of agreement on mastery status was 91 percent and 93 percent. Since the examiner is in charge of the test, the rate of agreement among examiners is especially important. These rates in French and Spanish were 94 percent and 92 percent for all tests. The agreement on mastery status among French and Spanish examiners was 94 percent and 96 percent. (French interviewer C rated as reliably as the French examiners and has since been moved to examiner status.)

The results of the experiment in German were somewhat different. The more reliable German raters were the interviewers rather than the examiners. The interviewers agreed with each other 100 percent in both areas of interest in this study. They never varied from each other more than a "plus." The figures for the two examiners are lower; they agreed with each other generally at the rate of 89 percent and they agreed on mastery status at the rate of 87 percent.

The explanation for the difference probably lies in the history of the raters' association with each other. The German examiners never worked together but rather succeeded each other in the job with no overlap. The German interviewers, on the other hand, have provided consistency in testing for more than ten years.

If we can draw any general conclusions from this study, they would be these: at the very least, 84 percent of examinees would receive the same rating from two independent raters. Or, more realistically, in similarly stringent situations 94 percent of the examinees would receive the same scores from different French raters. Ninety-three percent would receive the same scores from different German raters. Agreement on mastery status would be 94 percent, 96 percent, and 93 percent.

There is further reason to believe that the rate of agreement is higher in practice. There is no problem with lack of acoustic fidelity in a face-to-face interview. Grades are never decided by one rater alone (as was done in this study) but rather by two raters in consultation. Further, in a live test situation, each member of the testing team can gather the evidence necessary for a sound judgment, whereas in the experiment each rater had to make do with someone else's testing technique.

Some of the most interesting and revealing tapes from this point of view were those that received a broad range of scores. Some of them involved difficult decisions of factor weighting, such as near native fluency and pronunciation against serious grammatical errors; or a vocabulary inventory and comprehension worthy of an educated speaker, but without structural control; or good use of difficult grammatical features, but a vocabulary liberally strewn with inappropriate anglicisms. Tests like these do not easily fit one definition, yet a decision in terms of ability to do a job must be made.

Appendix A

List of Participants

Vicente Arbelaez
William Van Buskirk
Monique Cossard
Susana Framiñán
Catherine Hanna
C. Cleland Harris
Pauie Horn
Isabel Lowery
Joann Meeks
Juan José Molina
Alain Mornu
Margarethe Plischke
Robert Salazar
Harlie Smith
Patricio Solís
Blanca Spencer
Marina Wille Stinson
Marie-Françoise Swanner
Jack Ulsh
Agustín Vilches
Allen I. Weinstein

Appendix B

French Ratings

Test Number	Rater					
	A	B	C	D	E	F
1	1.0	1.0	1.5	1.0	1.5	1.0
2	3.5	4.0	3.5	4.0	3.5	3.5
3	3.5	4.0	3.0	4.0	3.0	3.5
4	.5	.5	1.0	.5	1.0	.5
5	4.0	4.5	3.5	4.0	4.0	4.0
6	2.0	2.5	3.0	3.5	3.0	2.5
7	2.5	2.0	2.0	2.5	2.0	2.0
8	1.0	1.0	1.0	1.0	1.0	1.0
9	4.0	4.5	3.0	3.5	3.5	3.0
10	1.5	2.0	2.0	2.0	2.0	1.5
11	2.5	2.0	2.0	3.0	2.5	2.5
12	3.5	3.5	3.0	3.0	3.5	4.0
13	3.5	4.5	5.0	5.0	5.0	4.5
14	.5	.5	.5	1.5	.5	.5
15	3.5	2.0	3.0	3.0	3.0	2.5
16	4.0	4.0	4.0	4.5	4.0	4.5
17	.5	.5	1.0	1.0	.5	1.0
18	3.0	3.5	3.0	3.5	3.5	3.0
19	5.0	5.0	4.5	5.0	5.0	5.0
20	2.0	2.5	2.5	2.5	3.0	1.5
21	3.0	3.0	3.0	3.0	3.5	2.5
22	2.5	2.5	2.5	2.5	2.5	2.0
23	1.5	1.5	1.5	2.0	1.5	1.0
24	.5	.5	1.0	.5	1.0	.5
25	3.0	3.0	3.0	2.5	3.0	2.0
26	4.0	3.5	4.0	3.5	4.0	4.0
27	1.5	1.5	1.0	1.5	1.5	1.0
28	3.0	3.5	3.0	3.0	3.5	3.0
29	5.0	5.0	5.0	5.0	5.0	5.0
30	1.5	1.5	1.5	2.5	1.5	1.5
31	3.0	3.0	3.0	2.5	2.5	2.0
32	2.0	1.0	2.0	1.5	1.5	1.0
33	2.5	2.5	2.5	3.0	2.5	2.5
34	1.5	1.5	1.5	1.5	1.5	1.0
35	1.0	.5	1.0	.5	1.0	.5
36	3.5	3.5	3.0	3.0	3.5	3.0
37	4.0	4.0	4.0	4.0	3.5	4.0
38	5.0	5.0	5.0	4.5	5.0	5.0
39	3.0	2.5	3.0	3.0	3.5	3.0
40	1.5	1.5	2.0	2.5	2.5	1.5
41	1.0	1.0	1.5	2.0	1.5	1.0
42	.5	.0	1.0	.8*	.5	.5
43	1.5	1.0	2.0	1.5	1.5	1.5
44	2.5	2.0	2.5	2.5	2.5	2.0
45	4.0	3.5	4.0	4.0	3.5	4.0
46	3.0	2.5	2.5	3.0	3.5	2.5
47	2.5	2.5	2.5	3.0	2.5	3.0
48	4.5	4.5	4.0	5.0	4.5	4.5
49	1.5	2.0	1.5	2.5	1.5	1.0
50	2.0	2.5	2.0	2.5	3.0	2.0

*Investigator supplied data (average score).

German Ratings

Test Number	Rater			
	A	B	C	D
1	1.0	1.5	.5	1.0
2	2.5	3.0	2.0	2.0
4	4.0	4.0	2.0	2.5
5	5.0	4.5	5.0	5.0
6	3.0	2.5	2.5*	3.0*
7	5.0	4.0	4.5	4.5
8	2.5	2.0	2.0	2.0
9	.5	.5	.5	.5
10	1.5	2.0	2.0	2.0
11	4.0	3.5	3.0	3.0
13	2.0	2.0	2.0	2.0
14	5.0	4.5	5.0	5.0
15	4.5	4.5	4.0	4.5
16	2.0	2.0	2.0	2.0
17	3.0	2.0	2.5	2.5
18	1.5	3.0	1.5	1.5
19	2.0	2.0	2.0	2.0
21	3.5	3.5	3.0	3.0
22	3.0	3.0	3.0	3.0
23	4.5	4.0	4.0	4.0
24	3.0	3.0	2.5	2.5
25	1.0	1.0	1.0	1.0
26	4.0	4.0	3.5	3.5
27	4.0	3.5	3.5	4.0
28	1.5	1.0	.5	1.0
29	3.0	1.5	2.5	2.0
30	1.5	2.0	1.5	1.5
31	3.5	2.0	2.0	2.0
32	3.0	2.5	3.0	3.0
33	.5	1.0	.5	.5
34	2.5	3.0	2.0	2.0
35	4.5	4.5	3.0	3.0
36	1.5	2.0	1.0	1.0
37	2.5	3.0	3.0	3.0
38	1.5	1.5	1.5	1.0
39	4.0	3.5	4.0	4.0
40	1.0	1.5	1.5	1.0
41	2.0	3.0	2.0	2.0
42	4.5	3.5	4.0	4.0
43	3.5	4.0	3.0	3.0
44	1.0	2.0	1.0	1.5
45	.5	.5	.5	.5
46	4.5	4.0	5.0	5.0
47	3.5	3.0	3.0	3.0
49	1.0	1.5	1.0	1.5
50	4.0	3.5	3.0	3.0

*Investigator supplied data (average score).

Spanish Ratings

Test Number	Rater A	B	C	D	E	F	G	H	I	J	K
1	1.5	1.5	1.5	1.0	1.0	2.0	1.0	1.5	1.5	1.5	2.0
2	2.0	1.5	1.5	1.5	1.5	2.0	2.0	2.0	1.5	1.5	2.0
3	1.5	1.5	1.5	1.5	1.5	1.5	1.0	1.5	1.5	2.0	2.0
4	4.5	5.0	4.5	4.5	4.5	4.0	5.0	4.0	4.5	4.5	4.5
5	2.5	2.5	2.0	2.0	2.0	2.5	2.0	2.0	2.0	2.0	3.0
6	4.5	5.0	4.0	3.5	5.0	3.5	5.0	3.5	5.0	5.0	4.5
7	1.0	1.5	1.0	.5	1.0	1.0	1.0	1.0	1.0	1.0	1.5
8	4.0	3.0	3.0	4.0	3.0	2.5	3.0	4.0	2.5	3.5	3.5
9	4.0	4.0	4.0	4.0	4.0	4.0	4.5	3.5	4.0	4.0	4.0
10	2.0	2.0	2.0	2.5	2.0	2.5	2.0	1.5	2.0	2.0	2.0
11	2.0	2.0	2.5	2.0	2.5	2.5	2.5	2.0	2.0	2.5	2.5
12	.5	.5	1.0	.5	.5	1.0	.5	.5	2.0	1.0	1.5
13	2.0	2.0	1.5	2.0	2.0	2.0	.5	2.5	1.0	2.0	2.5
14	1.0	1.0	1.0	1.0	1.0	2.0	1.5	1.0	1.5*	.5	2.0
15	2.0	2.0	2.0	2.5	2.5	2.0	2.0	2.5	1.5	2.5	2.5
16	3.0	3.5	3.5	3.5	3.5	3.5	4.0	3.0	3.5	4.0	4.5
17	2.0	2.0	2.0	2.0	2.0	2.5	1.5	2.0	1.5	2.0	3.0
18	3.5	3.0	3.0	3.5	3.5	3.5	3.5	3.5	2.0	3.5	3.5
19	.5	1.0	.5	.5	.5	.5	1.0	.5	.5	.5	.0
20	2.5	2.5	2.5	3.0	3.0	2.5	2.5	2.5	3.0	2.5	2.5
21	.5	.5	.5	.5	.5	.5	.5	.5	.5	.5	.5
22	1.0	.5	1.0	1.0	.5	1.5	.5	1.0	1.0	1.0	1.5
23	5.0	5.0	5.0	5.0	5.0	4.0	4.5	4.0	5.0	4.5	4.5
24	4.0	4.5	3.0	3.5	3.5	4.0	3.5	3.0	2.5	3.5	4.0
25	1.0	1.0	1.0	1.0	.5	1.0	2.0	1.0	1.5	1.0	1.5
26	3.0	3.0	3.5	3.5	3.5	3.0	3.0	3.5	2.5	3.5	3.5
27	3.5	3.5	3.5	3.5	4.5	4.5	3.0	3.5	3.5	4.0	4.5
28	2.5	3.0	2.5	2.5	4.0	3.0	2.5	3.5	3.5	4.0	4.0
29	2.5	2.5	2.0	3.0	2.5	2.5	2.5	3.0	2.5	2.5	3.0
31	3.0	2.0	2.0	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
32	4.0	4.5	4.0	3.5	3.0	3.5	4.5	3.0	4.5	4.5	3.0
33	1.5	1.5	1.5	2.0	1.5	2.0	1.5	1.5	1.5	1.5	2.0
34	1.0	1.0	1.0	1.0	1.0	2.0	1.5	1.0	1.5	1.5	2.0
35	2.5	2.5	2.5	2.0	2.5	2.5	1.5	2.0	2.5	1.5	3.0
36	4.5	5.0	4.5	4.5	4.5	4.0	4.0	4.5	4.5	4.5	4.5
37	1.5	1.5	1.5	1.5	1.5	2.5	2.0	1.5	1.5	1.5*	2.0
38	4.0	3.5	3.5	4.0	4.0	4.0	4.5	4.0	3.5	4.0	4.0
39	3.0	2.5	2.5	2.5	3.0	3.0	3.0	2.5	3.0	2.5	3.0
40	2.0	1.5	2.0	2.0	2.0	2.0	2.5	2.5	2.0	2.0	2.0
41	4.5	5.0	5.0	3.5	5.0	4.5	5.0	3.5	5.0	5.0	4.5
42	.5	.5	.5	1.0	.5	1.0	.5	.5	.5	.5	1.5
43	1.0	1.0	1.0	1.0	.5	1.5	1.0	1.0	1.0	1.0*	1.5
44	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.5	3.5	3.0	3.0
46	2.0	2.0	2.5	2.5	2.0	2.5	2.0	2.0	2.5	2.3*	2.5
47	1.5	1.0	1.0	2.0	1.5	1.5	1.5	1.5	1.0	1.5	2.0
48	2.5	2.0	2.0	2.0	2.0	2.5	4.0	2.0	1.5	3.5	2.0
49	2.0	2.0	2.0	2.0	2.5	2.0	2.5	2.0	2.0	2.5	2.0
50	4.5	4.5	4.0	4.0	4.0	4.0	4.5	3.5	4.0	4.5	4.5

*Investigator supplied data (average score).

References

Hambleton, R. K., and Novick, M. R. "Toward an Integration of Theory and Method for Criterion-Referenced Tests." Journal of Educational Measurement, 10 (1973): 159-70.

Millman, J. "Criterion-Referenced Measurement." In Evaluation in Education: Current Applications, edited by W. J. Popham. Berkeley, Calif.: McCutchan, 1974.

Swaminathan, H.; Hambleton, R. K.; and Algina, J. "Reliability of Criterion-Referenced Tests: Decision Theoretic Formulation." Journal of Educational Measurement, 11 (1974): 263-68.

Tollinger, S., and Paquette, F. A. The MLA Foreign Language Proficiency Tests for Teachers and Advanced Students. A Professional Evaluation and Recommendations for Test Development. New York: The Modern Language Association, 1966.

INDEPENDENT RATING IN ORAL PROFICIENCY INTERVIEWS

John Quiñones

Central Intelligence Agency

INDEPENDENT RATING IN ORAL PROFICIENCY INTERVIEWS

John ^uiñones

Background

Since 1972 the Language School of the Central Intelligence Agency has been using independent rating and averaging of testers' ratings to determine oral proficiency levels. In this paper I will discuss the development and use of a graphic rating scale that is used in conjunction with the verbal descriptions of the FSI Absolute Language Proficiency Ratings (Rice, 1959; FSI, 1963; Clark, 1972; Wilds, 1975).

The interview technique currently employed at the Language School is conceptually similar to the one developed at and used by the Foreign Service Institute of the U.S. Department of State. The two agencies differ, however, in three aspects. The testing team at the FSI consists of a native speaker of the language being tested and a scientific linguist familiar with the language. At the FSI, unlike the CIA, the S-rating is a combination of speaking and listening comprehension factors. The S-rating is always determined by averaging at the CIA (in languages in which there are at least two testers), while at the FSI there are several methods employed, including averaging when feasible.

Prior to 1972 the determination of proficiency levels in interview tests at the Language School was handled differently by different panels. In most cases one tester would suggest a rating and if the other member of the team disagreed, they would discuss the test until the discrepancy was resolved. In some cases the testers would vote on paper and if the ratings could be averaged (for example, one tester voted "2" and the other "3") they would combine the ratings. The resolution by discussion was sometimes time-consuming and occasionally led to unpleasant interpersonal confrontations, especially when one tester was inflexible in the interpretation of the level definitions.

Characteristics of the New Rating System

We felt the new rating system should contain at least the following features:

1. The system should allow for the differentiation of speaking and understanding since the Language School gives separate ratings for these skills.

2. The degrees of proficiency in each skill should be represented on the regular eleven-point scale, from 0 to 5, with "pluses" for levels 0 through 4.

3. The system should allow each tester the opportunity to contribute fully to the determiration of the final rating.

4. The system should facilitate immediate feedback to managers on the effectiveness of the testing program.

5. The system should permit immediate feedback to testers as to intertester reliability while at the same time decreasing or eliminating the possibility of interpersonal conflict.

6. The system should incorporate a graphic representation of the concepts of range and "plus."

7. The system should allow easy averaging of two or more ratings.

While most of the above desirable features did not require much clarification or discussion, we had to specify the notion of range in order to incorporate it into the graphic scale. I think that most practitioners of the FSI oral interview characterize the levels on the eleven-point scale as ranges. It is thus very common to hear, in the discussions that follow tests, statements such as "It's a low 3," or "It's a classical 3," or "It's a very high 2+ but not quite a 3." Because this notion is important in the assignment of levels, it is graphically represented on the scale that we developed (Figure 1). Testers can, therefore, make these finer judgments and have them count in the combined rating.

The notion of the "plus" was also made part of the graphic scale. In the current FSI system, all the numerical ratings except 5 may be modified by a "plus" to indicate that the examinee substantially exceeds the requirements for a level but fails to meet the requirements of the next higher level, especially in either grammar or vocabulary. The "plus" range is thus represented on the graphic scale as having a value of .60 to .99.

Rating of the Test

After the testers have finished the oral interview, they proceed to rate the examinee without consultation, using the rating sheets provided for this purpose (Figure 1). The independent judgment of each tester is expressed by drawing a line (---) across each rating scale (speaking and understanding) at the point he or she feels best indicates the examinee's overall proficiency in the skill. The testers are encouraged to make full use of the ranges in the scale since it is essential for the purpose of averaging the scores.

Determination of the Final Rating

After each tester has decided on his or her rating, the rating sheets, properly identified, are turned over to a testing aide for scoring. The combined rating for a given skill is determined by using a ruler marked in tenths. In cases in which the tester's mark coincides with a marking on the ruler, the lower tenth is always assigned. (The

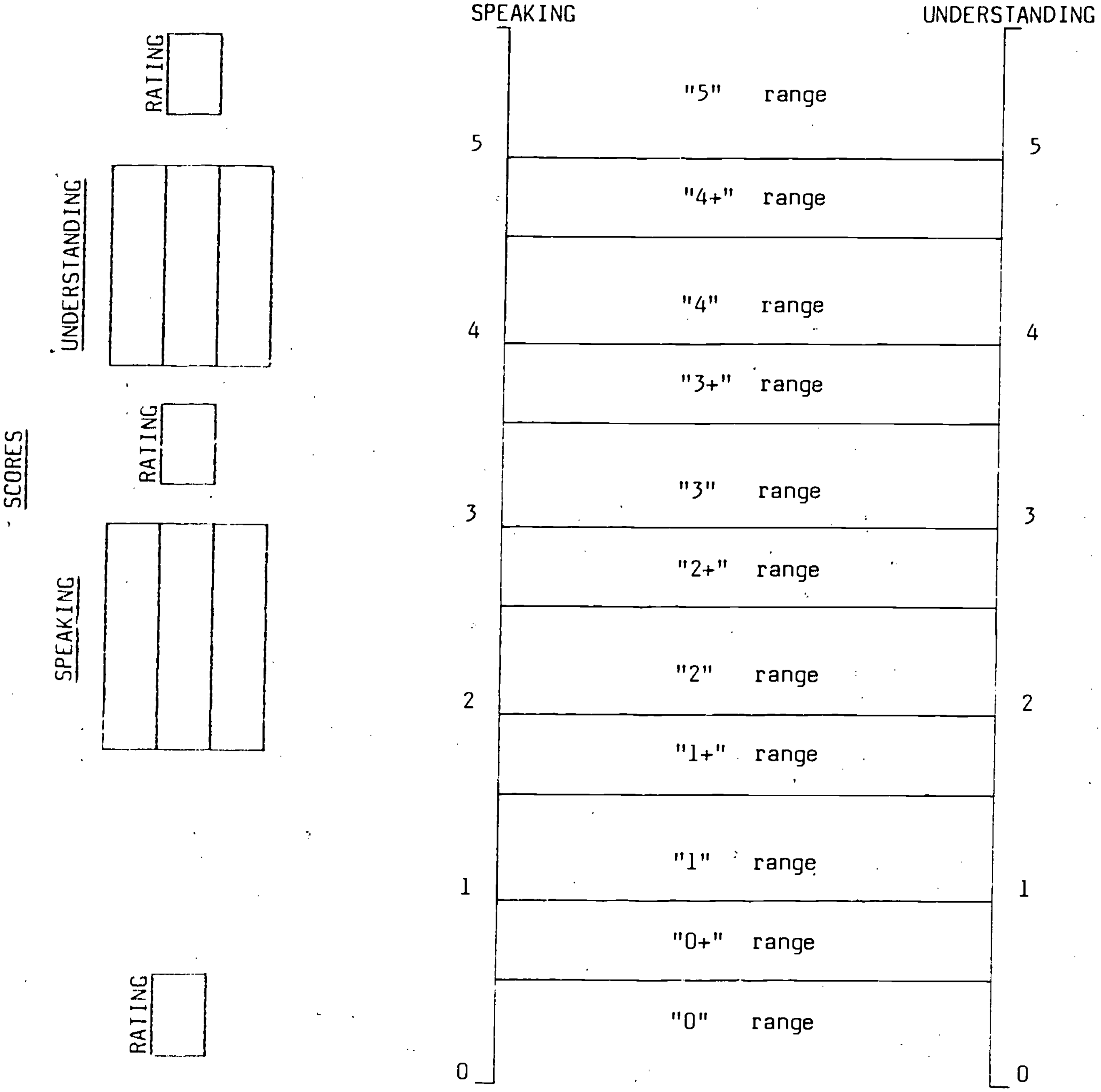
FIGURE 1

Language Proficiency Ra:ings
(Jral-Aural Skills)

Examinee ____________________ Language ______________

Examiner ____________________ Test Number ______________

Date ____________________



REMARKS:

rationale for this rule is that we believe that in general the consequences of overrating are more serious than the consequences of underrating.) Conversion tables are shown in Table 1.

This method of scoring permits not only the averaging of two scores but the averaging of any number of scores. The desirability of combined or averaged scores is supported by the fact that in oral interview tests (assuming that the testers are rigorously trained, as is presently the case in the Language School) the average is both a more reliable and a more accurate (valid) rating than the sole judgment of the best rater. This has been documented in studies on clinical judgment and decision making.

TABLE 1

Conversion Tables for
Language Proficiency Ratings

Overall Rating	Range of Scale Values for Single Judges	Range of Scale Values for Two Judges (Summed Ratings)	Range of Scale Values for Three Judges (Summed Ratings)
5	5.00	10.00	15.00
4+	4.60 - 4.99	9.20 - 9.99	13.80 - 14.99
4	4.00 - 4.59	8.00 - 9.19	12.00 - 13.79
3+	3.60 - 3.99	7.20 - 7.99	10.80 - 11.99
3	3.00 - 3.59	6.00 - 7.19	9.00 - 10.79
2+	2.60 - 2.99	5.20 - 5.99	7.80 - 8.99
2	2.00 - 2.59	4.00 - 5.19	6.00 - 7.79
1+	1.60 - 1.99	3.20 - 3.99	4.80 - 5.99
1	1.00 - 1.59	2.00 - 3.19	3.00 - 4.79
0+	0.60 - 0.99	1.20 - 1.99	1.80 - 2.99
0	0.00 - 0.59	0.00 - 1.19	0.00 - 1.79

	Range of Scale Values for Four Judges (Summed Ratings)	Range of Scale Values for Five Judges (Summed Ratings)	Range of Scale Values for Six Judges (Summed Ratings)
5	20.00	25.00	30.00
4+	18.40 - 19.99	23.00 - 24.99	27.60 - 29.99
4	16.00 - 18.39	20.00 - 22.99	24.00 - 27.59
3+	14.40 - 15.99	18.00 - 19.99	21.60 - 23.99
3	12.00 - 14.39	15.00 - 17.99	18.00 - 21.59
2+	10.40 - 11.99	13.00 - 14.99	15.60 - 17.99
2	8.00 - 10.39	10.00 - 12.99	12.00 - 15.59
1+	6.40 - 7.99	8.00 - 9.99	9.60 - 11.99
1	4.00 - 6.39	5.00 - 7.99	6.00 - 9.59
0+	2.40 - 3.99	3.00 - 4.99	3.00 - 5.99
0	0.00 - 2.39	0.00 - 2.99	0.00 - [illegible]

References

Clark, John L. D. *Foreign Language Testing: Theory and Practice.* Philadelphia: Center for Curriculum Development, 1972.

Foreign Service Institute. "Absolute Language Proficiency Ratings." Washington: Foreign Service Institute, 1963.

Jones, Randall L. "Testing Language Proficiency in the United States Government." In *Testing Language Proficiency*, edited by Randall L. Jones and Bernard Spolsky, pp. 1-7. Arlington, Va.: Center for Applied Linguistics, 1975.

Rice, Frank A. "The Foreign Service Institute Tests Language Proficiency." *Linguistic Reporter* 1 (1959): 2, 4.

Wilds, Claudia P. "The Oral Interview Test." In *Testing Language Proficiency*, edited by Randall L. Jones and Bernard Spolsky, pp. 29-38. Arlington, Va.: Center for Applied Linguistics, 1975.

THIRD RATING OF FSI INTERVIEWS

Pardee Lowe, Jr.

Central Intelligence Agency

THIRD RATING OF FSI INTERVIEWS[1]

Pardee Lowe, Jr.

This study has a long and school-wide genesis. It originated with an instructor's comment that Third Raters tend to place a candidate's speaking proficiency lower than do the interviewers who actually conduct the evaluation. Because this view is rather prevalent at the CIA Language School (LS) and, further, because the LS has maintained fairly complete Third Rater records dating back three years, this study seemed both feasible and desirable and, indeed, has proven enlightening, given the testing folklore to which we often unquestioningly subscribe.

Before turning to the methodology and the attendant results, the several types of raters need to be defined. At the LS, each language candidate's speaking proficiency is simultaneously, but independently, evaluated by two interviewers directly after a "live" oral interview. These interviewers will henceforth be referred to as Original Raters. Under certain conditions the opinions of one or more Third Raters will be called for. This might occur when there is a discrepancy between the ratings of the Original Raters, when the test score is disputed by the test candidate, or when the sample or the elicitation technique used to arrive at the sample strikes either of the Original Raters, their supervisor, or the chief of testing as unusual and worthy of closer scrutiny. For present purposes any rater who was not a member of the Original Rater team is regarded as a Third Rater. Thus, it is possible to speak of the first Third Rater, second Third Rater, and so forth, so long as these raters have listened to the same interview.

A second distinguishing characteristic of the Third Rater is that he or she is limited to evaluating an audio (only) tape recording of the interview. Furthermore, it should be emphasized that Third Rater data represent only the deviant cases (as defined above and which occurred from the beginning of 1975 through November 1977). The cases requiring three raters amount to less than 25 percent of all testing done in the languages reported. Security considerations preclude citing the actual percentage, which is considerably less than 25 percent. A more complete study would perforce include the vast majority of the evaluations in which there were no substantial disagreements between the Original Raters or any other reason to question the findings. The present study makes no pretense of being a thorough inter- and/or intrarater reliability and validity study.

[1] The present paper would not have been possible without the aid and support of the LS staff and instructors. I wish to express my gratitude to the LS instructor who raised the question about severity error in Third Raters' scoring and to Michael Gibbons and John Quiñones, who passed the question on and have read and given helpful suggestions on the present paper. Above all I wish to thank Robert J. Vincent, without whose help the statistical portions of the paper could not have been carried out.

Several arguments have been put forth in support of the hypothesis that a Third Rater tends to be a more severe judge than the Original Raters: (1) Third Raters listen to tapes; they are not present at the creation of the speech sample and therefore are not privy to the "richness" of the "live" performance. (2) A Third Rater has more time to concentrate on listening to the candidate's errors, for in the test itself the Original Raters have their hands full crchestrating the elicitation of the speech sample via an interactive question/answer interview. Although they may take notes during the process, much reliance upon memory goes into arriving at their final assessment. Thus, one might conclude that the Third Rater's increased opportunities to concentrate on the candidate's performance might unveil more errors, with the ccnsequence of a lower rating. (3) A Third Rater has the means to repeat (play back) any portion of the interview to check for errors, further increasing awareness of the number and types of errors. (4) A Third Rater may be asked to write out detailed comments and examples so that the test might be discussed more fully among the supervisor, the original testers, and the Third Raters. Again, the type and extent of these comments may lead one to predict a lower rating from the Third Rater.

A word or two at this point is in order on the matter of the ratings themselves. Figure 1 is an example of the language proficiency rating sheet on which each rater records his final assessment of the candidate's speaking prowess. For the candidate in question, Original Rater 1 scored the performance 2.8; Original Rater 2 was much more lenient (3.8). The large discrepancy led to a Third Rater being pressed into service (scoring the candidate 3.3, which, coincidentally, just happens to be the average of the scores set by the Original Raters). Each score is indicated along the speaking scale, as well as in the box to the left of the scale.

The official rating of each candidate is expressed as a range of proficiency (the eleven-point FSI rating scale), as depicted in Figure 1. In terms of the present data, Rater 1's score fell in the "2+" range, while Rater 2's score reached the "3+" range. Original Raters are considered to have arrived at the same proficiency evaluation if and only if each marks the same range, regardless of the actual numerical scale score. Since these raters' scores fell within different (indeed, discontinuous) ranges, a third rating seemed warranted. As noted, the third evaluation fell in the intermediate ("3") range.

Hypotheses

It should be clear at this point that each candidate's speech sample routinely receives two types of evaluations from each Original Rater: a numerical score and its corresponding FSI scale rating (encompassing a range of numerical scores).[2] Consequently it is quite possible for the

[2] For a fuller understanding of this process, see the John Quiñones paper on independent rating in this volume.

FIGURE 1

Language Proficiency Ratings
(Oral-Aural Skills)

Examinee ____________________ Language ______________

Examiner ____________________ Test Number ______________

Date ______________

Rating □

UNDERSTANDING

1.	2.	3.	Total

Rating 3

SPEAKING

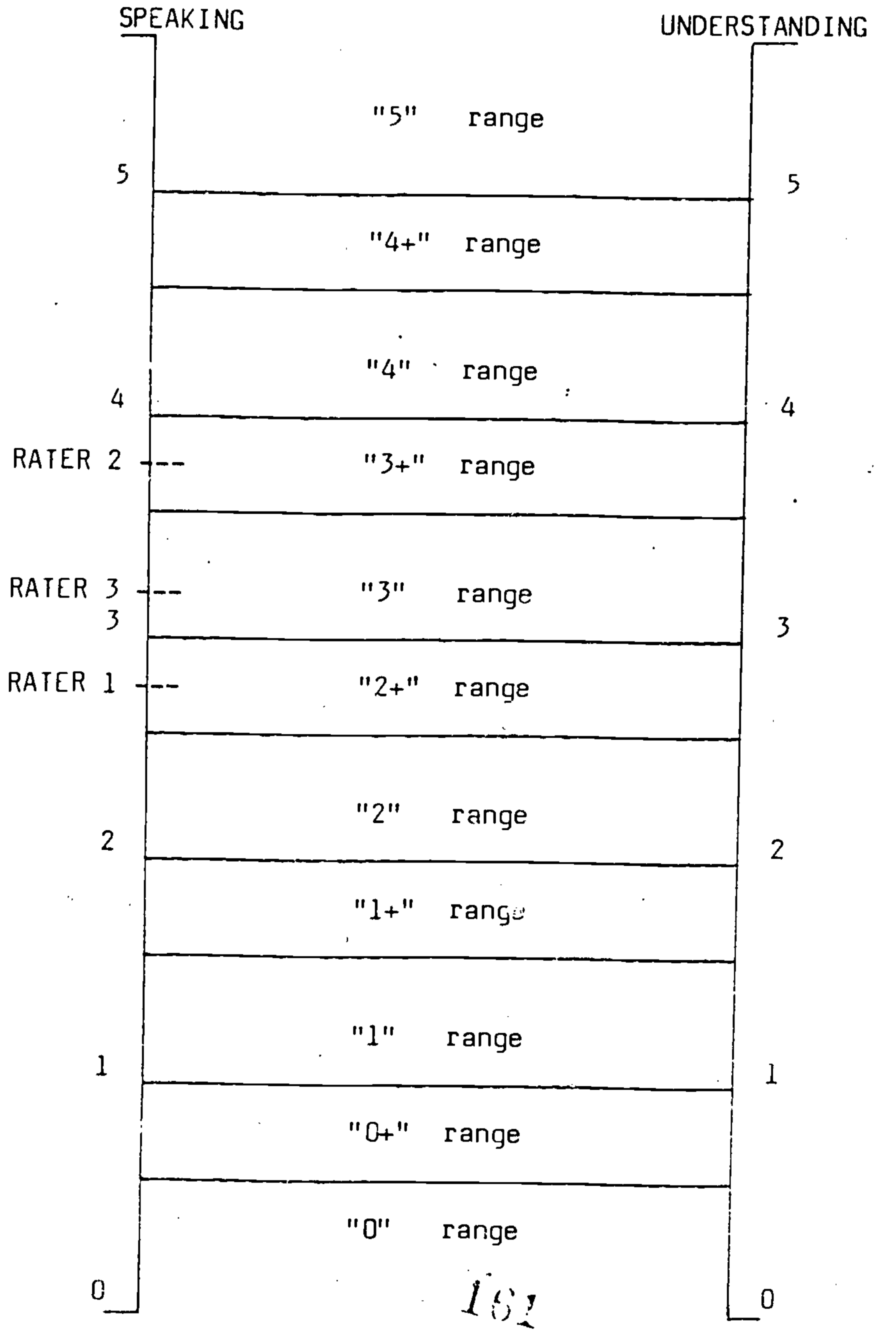
1. 2.8	2. 3.8	3. 3.3	Total 9.9

Rater 1: 2.8

RATER 2: 3.8

TOTAL 6.6

RATING □



REMARKS:

Original Raters' numerical scores not to agree, but if they fall within the same scale range the candidate will, in the last analysis, be judged similarly by both raters. Since the test of one type of criterion measure would not be complete without an evaluation of the other, two sets of hypotheses were established for testing, as follows:

Group 1 (FSI Scale Ratings)[3]

a. Third Rater evaluations fall in an FSI range above those of either of the Original Raters.

b. Third Rater evaluations fall in an FSI range below those of either of the Original Raters.

c. Third Rater evaluations fall in an FSI range intermediate to those of the Original Raters.

d. Third Rater evaluations are equal to at least one original rating.

Group 2 (FSI Numerical Ratings)[4]

e. The average Third Rater numerical rating is equal to the average of the Original Raters' numerical ratings.

f. The average Third Rater numerical rating is equal to the official numerical rating.

g. The average of the Original Raters' numerical rating is equal to the offical numerical rating.

Experimental Sample

LS records from 1975 through November 1977 were culled for each instance of a Third Rater evaluation. Sufficient numbers of such evaluations were found in French, Spanish, German, Russian, Chinese, Japanese, and Portuguese. In all, 163 examples were recorded, but for a variety of reasons, some of the analyses were restricted to a maximum of 149 cases.

[3] Analyses were restricted to data combined across all languages.

[4] Analyses were conducted on both individual-language and grouped data.

Procedure and Results

A series of chi-square analyses was conducted to test the Group 1 (FSI scale ratings) hypotheses. The chi-square test is ideally suited for testing whether a statistically significant difference exists between an observed number of events falling in each of several categories and an expected number based on the hypothesis that there are no systematic differences in the number of events in each of the categories. Table 1 summarizes the tests and attendant results associated with each of the Group 1 hypotheses.

Results of the first test indicate that the frequencies with which the ratings fell into the several categories (expressed as hypotheses a through d) were not equally distributed. In other words, there were systematic or nonchance differences in the manner in which the rating frequencies were distributed across the categories.

The second test addresses whether or not the frequency with which Third Raters grade below the Original Raters is comparable to the frequency with which they do not rate below. The answer, quite clearly, is that the number of times a Third Rater is more severe than both the Original Raters is more than offset by the number of times he is more lenient than at least one of them. As before, such differences are significantly nonchance.

Test 3 is concerned with the situation in which the Third Rater scores differently than both Original Raters. In other words, those cases wherein the Third Rater agreed with at least one of the Original Raters have been excluded from consideration. Here again there are significant differences in the frequencies among the categories, leading to the question posed in Test 4, which, paraphrased, reads: "When Third Raters' scores differ from those of at least one of the Original Raters, do Third Raters more often than not score lower?" The answer, in terms of FSI scale ratings, is most definitely no. Expressed another way, when Third Raters' scores do, in fact, differ from those of a least one of the Original Raters, there is as much chance that the Third Rater will score higher than at least one of the Original Raters as that he will score lower.

However, Test 5 indicates that significantly more Third Raters scored lower than both Original Raters than scored higher than both. The same was true in the comparison of the number of instances where Third Raters scored lower than both Original Raters to the situation where Third Raters' scores were in an FSI range intermediate of those of the Original Raters (Test 6).

Lastly, results from Test 7 show that for every instance where the Third Rater scored below both Original Raters, more than twice as often this rater scored higher than one of them.

TABLE 1

Tests of Group 1 Hypotheses
(FSI Scale Ratings)

Null Hypotheses:		Hypotheses* a	b	c	d	a+c (+d)
1. Equal numbers of cases fall in each category; N = 149.	OBSERVED	20	35	17	77	--
	EXPECTED	37.25	37.25	37.25	37.25	--
	Chi-square = 61.55; df = 3; p < .01; reject null hypothesis.					
2. Equal numbers of cases fall in each category; N = 149.	OBSERVED	--	35	--	--	114
	EXPECTED	--	74.5	--	--	74.5
	Chi-square = 41.89; df = 1; p < .01; reject null hypothesis.					
3. When Third Rater's score was different from those of both Original Raters, equal numbers of cases fall in each category; N = 72.	OBSERVED	20	35	17	--	--
	EXPECTED	24	24	24	--	--
	Chi-square = 7.75; df = 2; p < .05; reject null hypothesis.					
4. When Third Rater's score was different from those of both Original Raters, equal numbers of cases fall in each category; N = 72.	OBSERVED	--	35	--	--	37
	EXPECTED	--	36	--	--	36
	Chi-square = 0.06; df = 1; p > .05; accept null hypothesis.					
5. When Third Rater's score was different from those of both Original Raters, equal numbers of cases fall in each category; N = 55.	OBSERVED	20	35	--	--	--
	EXPECTED	27.5	27.5	--	--	--
	Chi-square = 4.09; df = 1; p < .05; reject null hypothesis.					
6. When Third Rater's score was intermediate or lower than those of both Original Raters, equal numbers of cases fall in each category; N = 52.	OBSERVED	--	35	17	--	--
	EXPECTED	--	26	26	--	--
	Chi-square = 6.23; df = 1; p < .05; reject null hypothesis.					
7. When Third Rater's score was lower than both or equaled at least one of Original Raters, equal numbers of cases fall in each category; N = 112.	OBSERVED	--	35	--	77	--
	EXPECTED	--	56	--	56	--
	Chi-square = 15.75; df = 1; p < .01; reject null hypothesis					

*Hypotheses: (a) Third Rater above both Original Raters.
(b) Third Rater below both Original Raters.
(c) Third Rater intermediate.
(d) Third Rater equal to at least one Original Rater.

In short, statistical analysis of FSI rating scale data does not support the contention that Third Raters are more severe assessors of speaking proficiency than are their Original Rater counterparts. To the contrary, Third Raters are as lenient or more so than at least one of the Original Raters better than 75 percent of the time, at least as far as FSI scale ratings are concerned.

Discussion thus far has been restricted to the FSI rating scale data. It was mentioned earlier that several hypotheses were generated concerning the comparability of the numerical ratings arrived at by the Original and Third Raters. To that end, an additional series of statistical analyses was conducted.

Table 2 summarizes the results on a language-by-language as well as an across-language basis. The sample size in these analyses totaled 163 (including the 149 reported earlier).

The earlier analyses dealt with the number of times an event occurred. The present situation has to do with actual scores, and, for that reason, another type of analysis is in order. To test for differences in numerical scores between and among the various types of raters, a statistical technique called a t-test was applied to the data. Like the chi-square test, the t-test determines whether the differences between groups (actually, pairs of groups) are statistically significant rather than attributable to chance variation.

Attention is directed first to the overall results at the bottom of Table 2. The average Third Rater numerical rating across all languages studied was 2.50 (a "2" on the FSI scale, since 2.6 would be required to reach the "2+" level). The rating arrived at by averaging the scores of the Original Raters was 2.62 (a "2+"). The difference between these numerical ratings was found by the t-test to be highly significant (and thus rejects hypothesis e).

A comparable analysis was concerned with the Third Rater/official rating relationship. Although the difference in average numerical ratings was found to be very significant (2.5 for the Third Raters; 2.38 for the official rating, rejecting hypothesis f), both sets of numerical ratings fell within the "2" rating scale.

The third overall comparison looked at the differences between average Original Rater numerical scores and the official ratings (expressed as numerical scores). Once again there were highly significant differences in favor of the Original Raters. Moreover, the corresponding FSI rating scores differed as well ("2+" vs. "2," respectively), rejecting hypothesis g.

A look at the individual language data reveals that what was true for the across-language data need not hold for any particular language. Although French and German Third and Original Raters disagreed beyond chance levels (Third Raters more severe) and similar, but not statistically significant differences were found for Spanish, Russian, and

TABLE 2

Tests of Group 2 Hypotheses
(FSI Numerical Ratings)

Languages	Average FSI Ratings	Third Rating	Average Original Rating	Third Rating	Official Rating	Average Original Rating	Official Rating
French	Numerical	2.56	2.77	2.56	2.47	2.77**	2.47**
	(Scale)	"2"	"2+"	"2"	"2"	"2+"	"2"
Spanish	Numerical	2.54	2.67	2.54	2.40*	2.67	2.40**
	(Scale)	"2"	"2+"	"2"	"2"	"2+"	"2"
German	Numerical	2.98	3.16*	2.98	2.94	3.16	2.94**
	(Scale)	"2+"	"3"	"2+"	"2+"	"3"	"2+"
Russian	Numerical	2.12	2.25	2.12	2.03	2.25	2.03**
	(Scale)	"2"	"2"	"2"	"2"	"2"	"2"
Chinese	Numerical	2.66	2.61	2.66	2.47	2.61	2.47
	(Scale)	"2+"	"2+"	"2+"	"2"	"2+"	"2"
Japanese	Numerical	2.27	2.12	2.27	2.00*	2.12	2.00
	(Scale)	"2"	"2"	"2"	"2"	"2"	"2"
Portuguese	Numerical	3.87	3.64	3.87	3.46**	3.64	3.46
	(Scale)	"3+"	"3"	"3+"	"3"	"3"	"3"
TOTAL	Numerical	2.50	2.62**	2.50	2.38**	2.62	2.38**
	(Scale)	"2"	"2+"	"2"	"2"	"2+"	"2"

* Probability of a difference this large due to chance less than .05.
**Probability of a difference this large due to chance less than .01.

Chinese, the opposite held true for Japanese and Portuguese (but the differences could easily be attributable to chance factors).

Third Rater vs. official rating comparisons revealed that Spanish, Japanese, and Portuguese Third Raters arrived at significantly higher numerical ratings than turned up in the official ratings. Each of the remaining languages followed suit, but the differences failed to reach conventional levels of significance.

Finally, average Original Rater numerical scores exceeded the official ratings in every case, with the differences for French, Spanish, German, and Russian highly significant.

With few exceptions, then, the numerical rating data indicate that there are highly significant differences among the official ratings (2.38), the Third Rater scores (2.50), and the average of the Original Raters (2.62). When these scores are converted to FSI scale ratings, however, both the official and Third Rater results are found to be more conservative ("2") than those for the Original Raters ("2+"). Therefore, the hypothesis that Third Raters grade more severely than the Original Raters is supported (in terms of both numerical scores and their equivalent scale ratings). This test contradicts the findings up to now. However, it must be remembered that arithmetic means are more influenced by a wide discrepancy in scores. This test, therefore, reflects variations in ratings. Since we deal at the LS in FSI levels (each of which comprises a range of scores), this test has the least significance for LS scores.

Third Raters tend to be more generous with their numerical scores than was reflected in the official ratings (although the corresponding scale ratings fell in the "2" range in both instances).

The import of this study for the LS and others who may opt to use an independent rating system with Third Raters is that, with properly trained personnel, severity error in Third Raters need be only a minor problem. Restricting our comments to the FSI rating analysis above, Third Raters were as lenient or more so than at least one of the Original Raters better than 75 percent of the time.

DETERMINING THE EFFECT OF UNCONTROLLED SOURCES OF ERROR IN A DIRECT TEST OF ORAL PROFICIENCY AND THE CAPABILITY OF THE PROCEDURE TO DETECT IMPROVEMENT FOLLOWING CLASSROOM INSTRUCTION

Karen A. Mullen

University of Louisville

DETERMINING THE EFFECT OF UNCONTROLLED SOURCES OF ERROR IN A DIRECT TEST OF ORAL PROFICIENCY AND THE CAPABILITY OF THE PROCEDURE TO DETECT IMPRUVEMENT FOLLOWING CLASSROOM INSTRUCTION

Karen A. Mullen

During the last few years, interest in direct testing of oral proficiency has grown. A number of research questions have been raised about the relationship between reliability and such variables as methods of scoring, length of testing time, number of interviewers, and number of interviews. In addition, questions have been posed about the relationship between direct and indirect tests of oral proficiency. I wish to present the results of a study undertaken in one ESL program to determine the answer to yet two more questions, one concerning the effect of uncontrolled sources of error in the procedure and the other involving the issue of whether the procedure can detect improvement in proficiency from one period to another. To allow for comparison between the FSI interview and the one described here, I will first note the context in which the study was conducted and then proceed to a description of the research design. I will then present the results and discuss the ways in which the oral interview may relate to indirect tests of oral proficiency.

At the time of this study, admission of a foreign student into an academic program at either the undergraduate or the graduate level at the University of Iowa was contingent upon academic eligibility and a TOEFL score of at least 480. The only exceptions to this were Vietnamese applicants, who were generally admitted without proof of eligibility or a TOEFL score report. Students whose TOEFL scores were between 480 and 550 or who had no scores to report were referred to our ESL program for further proficiency evaluation and recommendation to the ESL program if it seemed warranted.

As part of this evaluation, an examinee was interviewed by two instructors for fifteen to twenty minutes. One of the interviewers took the major responsibility for conducting the interview and the other listened, occasionally interjecting questions to clarify a misunderstanding or to move the conversation along in a natural and informal way. The intent was to make the interview as much like a real-life conversation as possible. At the beginning, the examinee was made to feel comfortable; talk usually centered around the weather, details of getting to the interview, country of origin, length of stay in the United States, and so forth. The interviewer then tried to find a broad topic on which the examinee could speak with some authority for a period of time. Usually examinees were asked to tell about their families, education, academic interests, goals, opinions, impressions, and attitudes. Interviewers were told not to modify their syntax or rate of speaking unless it became apparent that examinees did not understand. When this occurred, interviewers rephrased their questions and attempted to continue the conversations. If it was apparent that examinees were able to hold their own, every attempt was made to give them the opportunity to demonstrate their full ability to engage in communicative dialogue.

Following the interview, the two instructors rated the examinee on five scales of proficiency: listening comprehension, pronunciation, fluency, grammar, and overall proficiency. Each scale was represented by five continguous boxes of equal size, labeled poor, fair, good, above average, and excellent. Interviewers were instructed to put an "X" either inside the box or on the line between two boxes. These were later converted to numerical values (1 = poor, 2 = between poor and fair, 3 = fair, 4 = between fair and good, ... 9 = excellent). Interviewers consulted descriptions for the five levels of proficiency for each of the first four scales when determining the level. Overall proficiency was based on a subjective composite of the other four scales. The rating form and the skill-level descriptions are given in Appendix A.

To some degree, the procedure for assigning levels in this study differs from the FSI procedure. FSI interviewers are asked to make a global judgment first and then to fill out a five-scale checklist, with six intervals per scale. The global judgment on the FSI interview is not directly tied to any of the six intervals since the global judgment ranges from 0 to 5, with "pluses" in between. In this study, on the other hand, consideration of the four scales precedes overall judgment and the levels in each scale can be considered to be directly tied to the levels in the overall scale. Furthermore, unlike the case with the FSI interview, "vocabulary" is not one of the scales considered.

Interviewers in this study were ESL teachers who had had formal training in linguistics and language teaching and had taught ESL for at least one year. Because of the number of students to be interviewed and the time available for scheduling, the interviewers were randomly paired and assigned to interviews in two two-hour blocks, with a one-hour break between blocks, on each of three days. Examinees were randomly scheduled and assigned to the interviewing teams. No interviewer had ever met an examinee before the interview.

Following a semester of instruction, the subjects were interviewed again under the same format. To ensure that no instructional bias would be introduced in the second interview, interview teams were assigned to interview people who had not been students in their classes. These teams also interviewed new students who were referred to the program for evaluation and possible recommendation to ESL classes. As a result, they were not able to distinguish old students from new students.

The first objective of the study was to determine the best estimate of reliability for each of the testing periods. Reliability can be defined in a number of different ways. For the purpose of this study, I shall assume that in a situation in which a rater is given the task of estimating the magnitude of a specified characteristic for a given person in a single performance:

(1) the magnitude of the specified characteristic is constant; and

(2) the estimation of the specified characteristic by the rater consists of the constant magnitude just cited and an error of measurement that is due in part to the rater and in part to the conditions surrounding the measurement.

(1) is the "true score" and (2) is the observed score. For any number of raters under the same conditions, I further assume that:

(3) the true score of the person rated does not vary from rater to rater;

(4) the observed score of the person rated does vary from rater to rater; and

(5) the best estimate of that part of the score that varies from rater to rater is the mean error of measurement.

For any number of people to be evaluated, it is assumed that:

(6) the true scores will vary from person to person;

(7) the observed scores will also vary from person to person; and

(8) the variance of the observed scores is due in part to the variance in the true scores and in part to the variance in the mean error of measurement.

From (8) one may derive the equation:

(9) variance of observed scores = variance of true scores + variance of mean error.

If there were no variance in the mean error of measurement, the measurements would be 100 percent reliable. By the same token, the larger the variance in the mean error of measurement, the less reliable the measurements. Thus, the reliability of x raters is a ratio (where x is the number of raters):

$$(10)\quad \frac{\text{variance of true scores}}{\text{variance of true scores + variance of mean error of measurement}}$$

An analysis of variance provides an estimate of the variance of the mean error of measurement; in terms of the total variation, it is that part that is due to the variation within people. An analysis of variance will also provide an estimate of the variance of the observed scores, i.e., the denominator in (10); it is that part of the total variation that is due to the variation among people. These two estimates will be sufficient for determining the reliability of x measurements, where x is the number of raters:

$$\text{(A1)}\quad \frac{\text{average variation between people - average variation within people}}{\text{average variation between people}}$$

This estimate of reliability is biased since the average variation within people is affected by the number of people in the sample and the number of raters. Therefore, an adjustment must be made to produce an unbiased estimate:

$$\text{(A2)}\quad \frac{\text{average variation between people - m (average variation within people)}}{\text{average variation between people}}$$

$$\text{where m} = \frac{\text{(number of people) (number of raters - 1)}}{\text{(number of people) (number of raters - 1) - 2}}$$

In general, the unbiased reliability (A2) will be lower than the biased one (A1). The smaller the number of people in the sample or the smaller the number of raters, the larger the difference between A1 and A2. For example, were 2 raters employed, it would require a sample of 2,000 people for the difference between the two to be minimal. If the number of raters were increased to 3, a sample of about 1,000 people would be required for the difference to be minimal. If only 15 subjects were to be rated, it would require 135 raters for there to be a minimal difference between A1 and A2. Naturally, the smaller the number of people and the smaller the number of raters, the greater the difference between A1 and A2. Thus, for a small sample or a small number of raters or both, the unbiased reliability (A2) is the more appropriate statistic.

The variance of the average error of measurement, as mentioned in (2), includes the variance due to the main effect of raters as well as that due to uncontrolled errors. An analysis of variance that partitions the within-people variation into these two components makes it possible to further refine the estimate of observed-score variation due to uncontrolled errors of measurement. In this respect, we may consider that the within-people variation is composed of two subvariations; one is due to differences between raters and the other to errors not otherwise accounted for. We shall call this latter the residual variation. If we reconsider reliability in a way in which the effect of raters is not to be considered a part of the error of measurement, we then have a new definition patterned after that of A1 and A2.

$$\text{(B1)}\quad \frac{\text{average variation between people - average residual variation}}{\text{average variation between people}}$$

$$\text{(B2)}\quad \frac{\text{average variation between people - m (average residual variation)}}{\text{average variation between people}}$$

$$\text{where m} = \frac{\text{(number of people - 1)(number of raters - 1)}}{\text{(number of people - 1) (number of raters - 1) - 2}}$$

B1 is also known as Cronbach's alpha; it is a biased estimator. The addition of m into the B2 formula makes adjustments for sample size and the number of raters. With large samples, the two values of m in A2 and B2 will not differ appreciably. With small samples, m in A2 will be smaller in B2. In addition, if most of the within-people variation is due to differences between raters, the value subtracted is smaller. This, in combination with smaller m, will cause B2 to be greater than A2. B2 is the more appropriate if the effect of uncontrolled sources of error is the primary focus. B2 is also directly comparable to the Pearson product-moment correlation since neither depends on differences due to raters. In addition, following the suggestion of Ebel (1951), this formula is the more appropriate if decisions are based upon the average of the two ratings. The model upon which reliability is based is thus:

$$(11)\quad X_{ij} = \pi_i + \alpha_j + n_{ij}$$

Preliminary tests have shown that this model is appropriate. We may assume that the observed score is the sum of a true constant magnitude of the characteristic measured (π_i), the effect of rater (α_j), and the error of measurement (n_{ij}). Tukey's test for nonadditivity provides no evidence for the postulation of an interaction effect; that is, in all samples investigated, if one rater gives a higher rating than the other, he or she will consistently do so across all subjects. There is no evidence for suspecting rater A's giving higher scores to some subjects and lower scores to others while rater B does the opposite.

Table 1 shows the reliability of the mean of two measurements on each of the speaking proficiency scales for the nine samples of subjects evaluated in the first testing period and the fifteen samples from the second period. The chi-square tests in Table 2 show that, with the exception of the overall scale in the second testing period, the reliabilities of each testing period can be considered to be drawn from the same population ($p < .01$). The mean reliability for each testing period, determined by weighting each reliability according to the size of the sample from which it was calculated, is shown at the bottom of Table 1.

Nine of the rater pairs were the same for both testing periods. Paired t-tests indicate no significant ($p < .01$) difference in the mean reliabilities for the nine pairs in the two testing periods on any of the five scales (listening comprehension $t = .68$, pronunciation $t = .58$, fluency $t = .86$, grammar $t = .91$, overall $t = 1.02$). The correlation between the reliabilities of the first and second testing periods for the nine pairs are not positively correlated and so may be treated as independent samples. When the reliabilities of the six additional rater pairs in the second testing period are included, t-tests indicate no significant ($p < .01$) difference in the mean reliabilities for the two testing periods for all pairs (listening comprehension $t = .88$, pronunciation $t = 1.45$, fluency $t = 1.24$, grammar $t = 1.39$, overall $t = 1.61$). Since the mean reliabilities are not significantly different, the means of the mean reliabilities, determined again by weighting each reliability according to the size of the sample ($N = 115$, $N = 152$), are as follows: listening comprehension = .883, pronunciation = .781, fluency = .816, grammar = .796, and overall = .847.

TABLE 1

Reliability of the Mean of Two Measurements on Each of the Speaking Proficiency Scales for Rater Pairs for the Two Testing Periods

Pair	N		Listening		Pronunciation		Fluency		Grammar		Overall	
	First	Second	First	Second	First	Second	First	Second	First	Second	First	Secon
1	15	7	.923	.000	.773	.953	.759	.471	.853	.673	.893	.813
2	17	11	.833	.736	.868	.850	[illegible]	.000	.709	.000	.888	.000
3	10	15	.430	.926	.758	.909	.422	.953	.835	.913	.844	.968
4	17	10	.898	.890	.917	.747	.926	.693	.640	.675	.931	.720
5	25	12	.850	.419	.656	.724	.891	.801	.858	.738	.835	.713
6	14	15	.823	.851	.822	.872	.840	.868	.874	.844	.909	.854
7	7	7	.695	.781	1.000	.860	.889	.645	.973	.000	.925	.536
8	5	6	.980	.630	.362	.583	.600	.872	.870	.780	.785	.864
9	5	14	.897	.945	.241	.868	.879	.802	.864	.934	.818	.900
10	--	15	---	.890	---	.583	---	.847	---	.769	---	.818
11	--	7	---	1.000	---	.956	---	.906	---	.869	---	.978
12	--	7	---	.680	---	.155	---	.323	---	.766	---	.611
13	--	10	---	.874	---	.381	---	.844	---	.703	---	.844
14	--	10	---	.888	---	.715	---	.709	---	.786	---	.533
15	--	6	---	.720	---	.000	---	.372	---	.564	---	.242
Weighted Means	115	152	.851	.819	.775	.787	.851	.787	.826	.771	.885	

TABLE 2

Results of Chi-Square Tests on Reliabilities from the First Testing Period (number of interview pairs = 9) and the Second Testing Period (number of interview pairs = 15)

	Testing Period	
Scale	First (N = 9)	Second (N = 15)
Listening	9.54	25.50
Pronunciation	13.13	22.62
Fluency	8.80	24.78
Grammar	8.17	22.75
Overall	2.77	37.74*

*Significant at p < .01.

Since estimates of population parameters are acceptably high, it appears that errors of measurement in the observed scores do not loom large. Thus, interest now focuses on the question of whether direct testing of speaking proficiency under the conditions described is capable of showing improvement in performance from one testing period to another. One hundred seven subjects were tested in both periods. Table 3 shows the standard deviations and mean scores on each of the scales for the two testing periods. There is no significant (p < .01) difference between raters on any of the scales. The reliabilities for this set of subjects are within range. Table 4 shows that the mean of the mean scores on each of the scales in the second testing period is significantly higher than it is in the first period (p < .05). The difference is about one-half a level for each scale.

TABLE 3

Results of T-Tests on Mean Performance as Measured by Two Raters on Five Scales of Speaking Proficiency for Two Testing Periods Four Months Apart (N = 107)

Scale	First Testing Period						Second Testing Period					
	Rater	Mean	S.D.	t	p	r	Rater	Mean	S.D.	t	p	r
Listening	1	5.94	1.74	-.77	.43	.852	1	6.56	1.63	-.52	.60	.869
	2	6.03	1.74				2	6.61	1.66			
Pronunciation	1	5.63	1.31	.56	.57	.776	1	6.17	1.21	1.61	.10	.827
	2	5.56	1.51				2	6.02	1.31			
Fluency	1	5.50	1.53	-.59	.55	.840	1	6.01	1.44	-1.24	.21	.846
	2	5.57	1.60				2	6.14	1.57			
Grammar	1	5.56	1.30	1.19	.23	.872	1	5.94	1.31	-.86	.39	.812
	2	5.46	1.37				2	6.03	1.25			
Overall	1	5.65	1.35	1.42	.15	.887	1	6.08	1.25	.00	1.00	.847
	2	5.53	1.46				2	6.08	1.27			

TABLE 4

Results of T-Tests on Mean Performance as Measured by Five Scales of Speaking Proficiency for Two Test Periods Four Months Apart (N = 107)

Scale	Test Period	S.D.	Mean	t	p (one-tailed)
Listening	1	1.74	5.99	-4.30	.000
	2	1.63	6.59		
Pronunciation	1	1.40	5.59	-4.53	.000
	2	1.25	6.09		
Fluency	1	1.56	5.54	-3.96	.000
	2	1.50	6.07		
Grammar	1	1.33	5.50	-4.21	.000
	2	1.28	5.99		
Overall	1	1.40	5.99	-4.37	.000
	2	1.26	6.08		

TOEFL scores for the two testing periods were available for 18 of the 107 subjects. Improvement as measured by direct testing is evident for these subjects on four of the five scales ($p < .05$), as indicated in Table 5. Such improvement is also evident on three of the five parts of TOEFL, the most relevant of which is the listening comprehension subtest. This subtest is considered to be a reasonably good predictor of oral proficiency, and one would expect improvement as measured by the direct test to also show up on the TOEFL subtest. However, since the latter requires the subject to read as well as listen, one might suspect that improvement in reading proficiency accounts for the difference in performance on the listening subtests for the two periods. In fact, the TOEFL reading subtest does not indicate significant improvement from the earlier testing period to the later one. Therefore, the change in performance on listening comprehension appears to be due to a real change in aural proficiency. The results from the listening comprehension scale of the interview corroborate this conclusion. Moreover, if it is true that the listening comprehension subtest is an indirect measure of other oral skills, such as pronunciation or fluency, one would expect improvement in pronunciation and fluency in the interview. This is the case.

Likewise, if TOEFL and the interview are two ways of measuring the same thing and if TOEFL shows greater control over grammar, one would expect the interview to reflect this. By the same token, were TOEFL to show no improvement in English structure, this would show up in the interview as well. However, it is clear that this is not the case. It seems that the interview is measuring some aspect of control over English structure that TOEFL is not, and vice versa. Given the fact that the TOEFL structure subtest gives subjects the opportunity to make grammatical judgments after thinking about the possible choices and the interview does not, it may be that the TOEFL structure subtest is a measure of passive control over English grammar and that the interview is a measure of active control. If there is a difference between the two, one would expect improvement to be less likely in the latter. This interpretation receives support from the present study and may serve to explain why the TOEFL structure subtest shows improvement and the grammar scale of the interview does not.

To examine this claim further, let us examine the relationship between these two types of knowledge. For those who have studied English as an academic subject in their home countries and have had very little opportunity to use and apply knowledge of the language in their day-to-day activities outside the classroom, passive control over the language will exceed active control. If the interview is to be considered a means of testing active control and TOEFL is a means of testing passive control, and if passive control is greater than active control, one would expect no high degree of correlation between TOEFL and the interview in the first testing period. However, after a period of language instruction in the language to be learned, and after a period of time in which the subject is forced to conduct most of his day-to-day activities in the second language, one would expect greater active control as well as greater passive control. Moreover, one would expect a higher correlation between

TABLE 5

Results of T-Tests on Mean Performance as Measured by Five Scales of Speaking Proficiency and Subtest and Composite Scores of TOEFL for a Paired Sample of 18 Subjects

	Scale	Test Period	Mean	N	t	p (one-tailed)	S.D.
INTERVIEW	Listening	1	6.05	18	-1.86	.04	.98
		2	6.50				.98
	Pronunciation	1	5.36	18	-2.19	.02	.89
		2	5.83				1.05
	Fluency	1	5.41	18	-2.18	.02	1.01
		2	5.77				.91
	Grammar	1	5.52	18	-1.23	.11	.86
		2	5.86				.85
	Overall	1	5.50	18	-2.30	.01	.82
		2	5.88				.70
TOEFL	Listening	1	37.88	18	-7.53	.00	5.50
		2	48.94				7.72
	English Structure	1	38.40	15	-3.63	.00	5.48
		2	42.46				5.99
	Vocabulary	1	39.80	15	-.79	.22	8.01
		2	41.26				5.75
	Reading	1	43.33	15	-1.54	.07	6.74
		2	45.66				5.76
	Writing	1	38.60	15	-2.03	.03	7.44
		2	43.00				5.90
	Composite	1	397.33	15	-6.89	.00	44.24
		2	447.61				44.52

the interview and TOEFL for the second testing period. This, indeed, turns out to be the case, as indicated in Tables 6 and 7. The former shows correlations between the interview and TOEFL that are not significantly greater than zero. This is the first testing period, before instruction. The latter shows the correlations between the interview and TOEFL after instruction.

In some cases the correlations are significantly greater than zero. The listening and grammar subscales of the interview and the overall scale correlate with the listening comprehension subtest of TOEFL at a level greater than zero. The greater gains in TOEFL were those on the listening and structure subtests. We see that in contrast to the first testing period, in which the group was most homogeneous on these two scales and at the lower end, the reverse is true in the second testing period. If passive control over structure has increased, one would expect a concomitant increase in active control over that demonstrated in the first testing period. This should be related to levels of proficiency as demonstrated by the interview. Indeed, in the second testing period the listening and grammar scales of the interview are correlated with the TOEFL listening subtest at a level greater than zero. However, only about 20-25 percent of the variance of the two tests overlaps, indicating that the two tests are measuring independent aspects of listening comprehension as well.

The vocabulary subtest of TOEFL also correlates at a level greater than zero on four of the five scales of the interview. Vocabulary scores do not show a significant improvement from one testing period to the other, but because of an improvement in pronunciation skills, pronunciation scores correlate very highly with the vocabulary subtest scores for the second test period. It appears that passive control over the lexicon is not very different from one period to another but active control is. Words are more than visually recognized; they are now articulated more precisely as they are spoken. At the same time, recognition of words in the flow of speech has improved, as evidenced by the change in listening comprehension scores. Thus, the higher correlation between listening comprehension scores in the interview and the vocabulary scores in TOEFL is an indication of greater active control over the lexicon.

No improvement on the grammar scale of the interview is evidenced, nor is improvement on the vocabulary subtest of the TOEFL. Yet a correlation greater than zero exists between these two scales in the second testing period but not in the first. I have no explanation for this fact. Neither can I offer an explanation for the nonzero correlation between the pronunciation scale of the interview and the writing ability subtest of TOEFL.

In general, this study suggests that the correlation between TOEFL subscores and the interview scores will be nonexistent when there is little active control of English. As active control of the language improves, the correlation between TOEFL subscores and the interview scores

TABLE 6

Correlation of TOEFL with Five Scales of Speaking Proficiency for First Testing Period (N = 18)

	TOEFL					
Interview Scale	LC (N=18)	ES (N=15)	Voc (N=15)	Rdg (N=15)	WA (N=15)	Composite (N=18)
Listening	.21	.37	.02	.00	.23	.24
Pronunciation	.27	.04	-.07	.10	.22	.14
Fluency	.25	.23	.12	-.10	.22	.21
Grammar	.09	.00	.07	-.03	.02	.06
Overall	.26	.12	-.06	-.04	.11	.10

TABLE 7

Correlation of TOEFL with Five Scales of Speaking Proficiency for Second Testing Period (N = 18)

	TOEFL					
Interview Scale	LC (N=18)	ES (N=15)	Voc (N=15)	Rdg (N=15)	WA (N=15)	Composite (N=18)
Listening	.45*	-.06	.46*	-.08	-.27	.28
Pronunciation	.19	.29	.70**	-.24	.51*	.26
Fluency	.36	.24	.42	.05	.14	.39
Grammar	.47*	.40	.46*	.08	.32	.48*
Overall	.54**	.09	.44*	.13	.14	.43*

* p < .05.
** p < .01.

becomes stronger. The mean TOEFL score for these subjects is below the level one would judge necessary for full participation in an English-speaking class. Though the data are not available here to verify the prediction, one would expect that as speaking proficiency as measured by the interview continued to improve, a nonzero correlation between the grammar scale of the interview and the structure subtest of the TOEFL would begin to surface. This bears further investigation.

The major conclusions to be drawn from this study are that direct testing of speaking proficiency under the conditions described is a fairly reliable procedure and that the interview cannot be expected to correlate with subtests of the TOEFL when proficiency is low and passive control exceeds active control. As the difference between active and passive control diminishes, the correlation between TOEFL and a direct oral proficiency test can be expected to be greater than zero. The claim is that where performance on direct tests of oral proficiency is at a high level, TOEFL will tell us that, and where performance on a direct test of oral proficiency is low, there is no way to tell if it is due to a general lack of knowledge about the language or lack of skill in speaking and listening.

Appendix A

Interview Evaluation

Name ______________________ Date __________ __________

Evaluator ______________________

	Poor	Fair	Good	Above Average	Excellent
Comprehension					
Pronunciation					
Fluency					
Grammar					

Overall Oral Proficiency					

Guidelines for Evaluation of Interviews

Comprehension

Excellent: Appears to understand everything without difficulty.
Very Good: Understands at nearly normal speed; occasional repetition necessary.
Good: Understands at slower-than-normal speed; frequent repetition necessary.
Fair: Great difficulty following questions and answers.
Poor: Cannot be said to understand even simple conversation.

Pronunciation

Excellent: Has few traces of foreign accent.
Very Good: Always intelligible, though definite accent present.
Good: Concentrated listening is necessary; errors cause occasional misunderstanding.
Fair: Very hard to understand; repetition frequently necessary.
Poor: Speech virtually unintelligible.

Fluency

Excellent: Speech as fluent and effortless as that of a native.
Very Good: Fluency slightly affected by language problems.
Good: Fluency rather strongly affected by language problems.
Fair: Usually hesitant; forced into silence by language problems.
Poor: Halting and fragmentary speech; conversation impossible.

Grammar

Excellent: Few, if any, noticeable errors of grammar or word order.
Very Good: Occasional grammatical and/or word-order errors.
Good: Frequent grammar and word-order errors that obscure meaning.
Fair: Comprehension difficult; frequent rephrasing; uses basic patterns.
Poor: Severe errors in grammar and word order.

References

Clark, John L. D. "Theoretical and Technical Considerations in Oral Proficiency Testing." In *Testing Language Proficiency*, edited by Randall L. Jones and Bernard Spolsky, pp. 10-28. Arlington, Va.: Center for Applied Linguistics, 1975.

Ebel, Robert L. "Estimation of the Reliability of Ratings." *Psychometrika* 16 (1951): 407-24.

Hinofotis, Frances B. "Cloze Testing as a Substitute for Oral Interviews." Paper presented at the Preconference Workshop on Cloze Testing, TESOL Conference, Miami, Fla., 1977.

Hoyt, Cyril J. "Test Reliability Estimated by Analysis of Variance." *Psychometrika* 6 (1941): 153-60.

Mullen, Karen A. "Rater Reliability and Oral Proficiency Evaluations." In *Occasional Papers on Linguistics: Proceedings of the First International Conference on Frontiers in Language Proficiency and Dominance Testing, Carbondale, Illinois, 1977*, edited by James Redden, pp. 133-42. Carbondale: Department of Linguistics, Southern Illinois University, 1977.

Wilds, Claudia P. "The Oral Interview Test." In *Testing Language Proficiency*, edited by Randall L. Jones and Bernard Spolsky, pp. 29-44. Arlington, Va.: Center for Applied Linguistics, 1975.

Winer, B. J. *Statistical Principles in Experimental Design*. 2d ed. New York: McGraw-Hill, 1971.

RELIABILITY AND VALIDITY OF LANGUAGE ASPECTS CONTRIBUTING TO ORAL PROFICIENCY OF PROSPECTIVE TEACHERS OF GERMAN

Ray T. Clifford

Central Intelligence Agency

RELIABILITY AND VALIDITY OF LANGUAGE ASPECTS CONTRIBUTING TO ORAL PROFICIENCY OF PROSPECTIVE TEACHERS OF GERMAN

Ray T. Clifford

Introduction

It has long been accepted as axiomatic that foreign language teachers must be proficient in the languages they teach. Axelrod (1966, p. 7) defines the "excellent foreign language teacher" as one who, along with other skills," . . . speaks the language intelligibly and with adequate command of vocabulary and syntax." The MLA statement of "Qualifications for Secondary School Teachers of Modern Foreign Languages" (1955, pp. 46-47), hereafter referred to as the MLA Teacher Qualifications Statement, was reaffirmed in 1966 (Paquette, p. 373). It describes three levels of oral proficiency and includes a description of the situations where these skills are to be demonstrated:

> Minimal--The ability to talk on prepared topics (e.g., for classroom situations) without obvious faltering, and to use the common expressions needed for getting around in the foreign country, speaking with a pronunciation readily understandable to a native.
>
> Good--The ability to talk with a native without making glaring mistakes, and with a command of vocabulary and syntax sufficient to express one's thoughts in sustained conversation. This implies speech at normal speed with good pronunciation and intonation.
>
> Superior--The ability to approximate native speech in vocabulary, intonation, and pronunciation (e.g., the ability to exchange ideas and to be at ease in social situations).
>
> Test--For the present, this ability has to be tested by interview or by a recorded set of questions with a blank disc or tape for recording answers.

It is interesting to note that this statement, published long before the debate over linguistic and communicative competence developed, recognized a combination of both linguistic and communicative skills. Much of the discussion of "communicative competence" is directed toward students and does not include the linguistic skills that would be expected of a teacher who provides a model of the target language for his students. Likewise, it can be assumed that teachers must have a communicative competence beyond simple linguistic competence if they are to teach others to communicate effectively. Therefore, the term "language proficiency" will be used in this study in its broadest meaning, encompassing both linguistic and communicative skills.

At this point, only two generally accepted methods of testing oral proficiency in foreign languages have been developed: the speaking portion of the MLA Cooperative Foreign Language Proficiency Test and the FSI interview procedure. Of these two procedures, only the MLA test has been used in assessing the language skills of pre- and inservice teachers. Clark (1975) contends that an interview could be used to test teachers and, accor( ) to him, it would be a more direct, and therefore a more valid, measu.e of language proficiency than the generally used MLA tests. The authors of the MLA Teacher Qualifications Statement quoted above also considered an interview as a possible mode of oral proficiency assessment. Although oral interviews are used by several government agencies, including the Foreign Service Institute, the CIA, the Peace Corps, and the Civil Service Commission (Wilds, 1975; Lowe, 1976), these techniques have not been widely used in or specially adapted to testing the language proficiency of teachers.

To be useful, a language proficienc test must be both valid and reliable. The development (a proficiency interview for teachers would provide two independently constructed tests of oral proficiency, which would allow inferences about the concurrent and construct validity of those measures and about the relative reliability of an indirect measure of oral proficiency as compared to a direct interview situation.

Research Problem

This study developed an oral interview procedure for testing prospective teachers of German by adapting the established FSI interview procedures used by governm.:.t agencies to more closely parallel the MLA proficiency definitions. It then compared this Teacher Oral Proficiency assessment procedure with the only existing standardized test of foreign language competence for teachers that includes the testing of speaking skills: the MLA Cooperative Foreign Language Proficiency Test.

The study also examined the concept of language aspects thought to contribute to oral proficiency. Both the FSI and MLA testing procedures identify the same four factors as contributing to oral language proficiency: structure or grammar, vocabulary, pronunciation, and fluency. These four aspects of oral language are also included in the language testing models proposed by Lado (1961), Cooper (1968), Carroll (1968), Harris (1969), and Valette (1971). However, both the MLA and FSI scoring procedures yield onl: overall scores, thus masking the contribution of the individual scores used in arriving at a total score.

A total test score implies a homogeneity of subcategories within the test. If, on the other hand, the scoring subdivisions used are independently valid, each should receive a separate score. One of the conclusions reached by the Minnesota Council of Teachers of Foreign Languages Working Committee on Teacher Certification in 1976 was that teachers should be at least minimally proficient in each of these areas and not just very good in any one of the language aspects being considered. Thus, if structure, vocabulary, pronunciation, and fluency do contribute independently to general oral language proficiency, scores should be computed separately for each factor--both for providing descriptive levels of proficiency with diagnostic value and for setting minimum levels for the certification of teachers.

No empirical evidence has been produced that points toward the validity of these contributing factors to oral language proficiency, but a statistical procedure suggested by Campbell and Fiske (1967) seems ideally suited to providing such evidence. Referred to as "convergent and discriminant validation," this procedure requires not only that indicators of a hypothesized factor converge (i.e., show high positive correlation with each other), but that they also be distinguishable from each other. In statistical terms, this means that the indicators of each hypothesized factor correlate more highly with other indicators of the same factor than with indicators of other factors.

In summary, the main questions investigated in this study may be briefly stated as follows:

1. Is it possible to structure a valid and reliable oral interview and rating procedure for directly assessing the oral language proficiency of prospective teachers of German?

2. What is the correlation between oral proficiency scores obtained from the "direct" assessment procedure and scores from the speaking test of the MLA Cooperative Foreign Language Proficiency Test?

3. What are the interrater, intrarater, and test-retest reliabilities for the speaking portion of the German MLA Cooperative Foreign Language Proficiency Test and for the oral interview procedure in the same situation?

4. Do measures of the same aspects of oral language, arrived at by these different testing procedures, correlate more highly with each other than they do with other language aspects measured by either procedure?

Procedures and Instrumentation

The target population for this study was prospective teachers of German enrolled at the University of Minnesota. Because of the limited number of students applying for admission to the College of Education during any one school year, the sample size was increased by including in the investigation all students who, in terms of language courses completed, were eligible to apply for admission during the 1975-76 school year, whether they actually did apply or not. In all, fifty students were contacted and forty-seven participated in the study.

The proficiency test used in this study was the Speaking Test, Form HC, of the MLA Cooperative Foreign Language Proficiency Tests: German, formerly called MLA Foreign Language Proficiency Test for Teachers and Advanced Students: German (Buros, 1972). In this test of oral proficiency in German, students' responses to prerecorded and visual stimuli are recorded on audio tape for later scoring. The test lasts fifteen minutes and is divided into three parts. In Part A the examinee hears twenty recorded statements that he is to repeat. He is then scored on the correctness of his pronunciation on two selected phonetic elements in each of the last fifteen statements presented.

Part B contains a printed selection that the examinee reads first to himself and then aloud. His pronunciation is again rated, on twenty selected phonetic features of the language, and his reading fluency is also rated, according to a five-point scale ranging from failure to convey the meaning of the passage to performance like a native who reads well.

In Part C the examinee is asked to describe orally a picture or a series of pictures. He is given three opportunities to respond ranging in duration from forty-five to ninety seconds per picture or series of pictures. The examinee's performance is rated separately for each of the three picture situations in each of the areas of vocabulary, pronunciation, structure, and fluency. The rating scales are specific to each area, but all are rated according to a five-point scale ranging from inadequate to native performance. The resulting twelve ratings are totaled to arrive at the examinee's score on Part C.

The interview and rating procedure specifically designed to test prospective teachers of German was named the Teacher Oral Proficiency (TOP) interview. It was developed by combining the various proficiency rating scales available into one general rating scheme that could be used in an interview situation to test the oral language proficiency of teachers. For this purpose a separate six-by-six matrix was developed for each of the language aspects of grammar, vocabulary, pronunciation, and fluency. One dimension of each matrix was divided into six proficiency levels, designated 0 to 5, and the other dimension was divided into categories according to the six available rating scales: the MLA Teacher Qualifications Statement, the rating scale from the MLA speaking proficiency test, the general FSI proficiency descriptions, the FSI grid of "Factors in Speaking Proficiency," the FSI supplementary proficiency descriptions, and the CIA supplementary rating criteria.

Not all six rating scales described each skill area of grammar, vocabulary, pronunciation, and fluency at each proficiency level, but each level was described by at least one rating scale. The matrices for grammar, vocabulary, pronunciation, and fluency were then presented to a "Second Languages and Cultures Education" seminar at the University of Minnesota, where graduate students and faculty members eliminated redundant proficiency descriptions in the rating scales. This left a matrix of the unique contributions provided by each rating scale in describing each aspect of oral proficiency at each level of proficiency. These four matrices were then collapsed to form one rating grid with separate rating scales for each language aspect.

The combined rating grid was used both as a framework for structuring TOP interviews and as a rating scale for evaluating performance in those interviews. A TOP interview lasts fifteen to thirty minutes and is conducted in much the same way as an FSI interview. It may be conducted by one or two interviewers, who begin the interview with simple questions about general topics and then broaden the discussion as far as the language skills of the interviewee permit. When it is evident that the interviewee has been pushed beyond his highest level of performance, the discussion is returned to more general topics before the interview is ended, so the interviewee will not perceive the experience as negative or frustrating. Ratings are assigned separately for the interviewee's performance in the areas of grammar, vocabulary, pronunciation, and fluency.

The MLA speaking test and TOP interviews were administered twice each to the forty-seven students participating in the study. All tests and interviews were recorded on cassette tapes for later scoring by the author and three other raters, all native speakers of German, trained by him. Tapes from the first administration of the MLA speaking test were scored first, then the tapes from the second MLA test administration. This was followed by a rescoring of the tapes from the first MLA test administration. The same procedure was followed in rating the taped TOP interviews, so each rater supplied three interview ratings and three MLA speaking test scores for each student.

These scores and ratings were then correlated to determine the reliability and validity of the MLA and TOP measures of oral language proficiency in German. Different computational procedures were used depending on the question to be investigated. Pearson product-moment correlations were calculated to estimate validity, while intraclass correlations were used to estimate the respective reliability of both testing procedures. Convergent and discriminant validation criteria as established by Campbell and Fiske (1967) were applied as a test of the construct validity of the language aspects: grammar, vocabulary, pronunciation, and fluency.

Several limitations are evident in this study. A major limitation results from the relaxed criterion used in selecting the sample of students to be tested. A sufficient number of students was tested to allow meaningful inferences about the theoretical relationship under study; however, the tested sample is one step removed from a truly representative sample of prospective teachers of German. Another limitation is that all oral interviews were conducted by the same interviewer, making it impossible to measure or infer how much variance in students' scores might be caused by the interaction of interviewer and interviewee characteristics.

A third limitation is that, in an examination of concurrent validity, no one measure of proficiency can be assumed as the standard against which the other may be judged. Thus, a low correlation "casts doubt on both measures, presumably equally" (Cronbach, 1971, p. 466).

Results of the Study

From a subjective viewpoint, this attempt at developing and using an interview procedure to test prospective teachers of German was a success. The modified rating scale served well as an underlying structure for conducting the interviews, and the raters experienced little difficulty in rating interviewees' performance according to that scale. Empirically, the results were also favorable.

A. Concurrent validity

Concurrent validity of the MLA test and TOP interviews was estimated by computing Pearson product-moment correlations. Total scores from the interviews correlated .834 with total MLA speaking test scores and .864 with global ratings assigned in Part C of the MLA speaking test.

B. Reliability

All reliability coefficients were computed using intraclass correlational formulas, which--unlike product-moment correlations--treat differences among the means of the correlated scores as error variance.

1. Interrater reliability

For both testing procedures, ratings of individual language aspects were less reliable than the sums of those ratings. The intraclass, interrater reliability of total scores on the MLA test was .818, while for Part C it was found to be .829. The intraclass, interrater reliability of sums of ratings from TOP interviews was .827. The interrater reliability of the language aspect ratings on both testing procedures is summarized in Table 1.

2. Intrarater reliability

The mean intraclass, intrarater reliability coefficients for total scores followed the same pattern found with interrater reliability. The mean intrarater reliability of Part C of the MLA speaking test was found to be .911, which is slightly larger than the mean intrarater reliability of .897 found for total MLA speaking test scores. The mean intrarater reliability of sums of ratings on the TOP interview was .930. The mean intraclass, intrarater reliability coefficients for language aspect ratings from both testing procedures are summarized in Table 2.

3. Test-retest reliability

The intraclass, test-retest reliability of total MLA speaking test scores and those for Part C of the MLA test were both .940, while the test-retest reliability of sums of ratings from TOP interviews was found to be .893. As Table 3 shows, the test-retest reliabilities of individual language aspects were lower when rated from the interviews than when rated from the MLA speaking test.

TABLE 1

Interrater Reliability of
Language Aspect Ratings

Language Aspect	Part C, MLA Speaking Test	TOP Interview
Grammar	.709	.719
Vocabulary	.770	.699
Pronunciation	.676	.690
Fluency	.801	.717

TABLE 2

Mean Intrarater Reliability of
Language Aspect Ratings

Language Aspect	Part C, MLA Speaking Test	TOP Interview
Grammar	.773	.903
Vocabulary	.853	.867
Pronunciation	.826	.836
Fluency	.857	.780

C. Construct validity of contributing language aspects

The mean scores of the language aspect ratings assigned students on the first administration of both the MLA test and TOP interview are given in Table 4.

It is interesting that the same relative ordering of mean scores on grammar, vocabulary, pronunciation, and fluency was found on both tests. Students were rated highest on pronunciation, followed in descending order by fluency, grammar, and vocabulary.

TABLE 3

Test-Retest Reliability of
Language Aspect Ratings

Language Aspect	MLA Speaking Test	TOP Interview
Grammar	.920	.859
Vocabulary	.885	.791
Pronunciation	.923	.881
Fluency	.908	.803

TABLE 4

Variables Examined for Construct Validity
of Contributing Factors

(N = 47 for all variables)

Test	Language Aspect	Mean	Standard Deviation
MLA	Grammar	7.82	1.78
MLA	Vocabulary	7.40	2.11
MLA	Pronunciation	8.76	2.04
MLA	Fluency	7.88	2.13
TOP	Grammar	2.39	0.69
TOP	Vocabulary	2.25	0.62
TOP	Pronunciation	2.64	0.63
TOP	Fluency	2.53	0.71

A correlation matrix of the variables in Table 4 is found in Table 5. This matrix of product-moment correlations was used to examine the ratings of grammar, vocabulary, pronunciation, and fluency for construct validity, according to the criteria for convergent and discriminant validation. The three essential criteria are:

1. All correlation coefficients in the validity diagonal of the multitrait, multimethod triangle should be statistically significant and sufficiently large to indicate convergent validity.

2. Each trait correlation coefficient in the validity diagonal should exceed in magnitude the correlations of that trait with other traits measured by a different method.

3. Each trait correlation coefficient in the validity diagonal should exceed in magnitude the correlations of that trait with other traits measured by the same method.

The validity coefficients in Table 5 have been underlined. The conditions of criterion number 1 above were met by those correlations found on the validity diagonal of the matrix. The conditions of criterion number 2 were met for the language aspects of pronunciation and fluency, but, because of a high correlation between TOP grammar ratings and MLA vocabulary ratings, they were not met for the language aspects of grammar and vocabulary. The conditions specified by criterion 3 were not consistently met by any of the validity correlations. The multitrait, multimethod correlation matrix in Table 5 gave some indication of convergent and discriminant validation, but because of apparent method variance introduced by the particular testing procedure used, none of the language aspects met the conditions of criterion number 3. Therefore, validation of the language aspects hypothesized as contributing to oral language proficiency was not achieved using this multimethod matrix.

A multitrait, multirating matrix of the correlations between average first and second ratings of the same test administration showed different results. The resulting matrix for the TOP interview is shown in Table 6, and the matrix for the MLA test is in Table 7.

Correlating the mean scores assigned students on the hypothesized language aspects on the first and second ratings of the same test administration for each procedure in effect controlled for error variance in the students' scores resulting from method variance, interrater variance, and trait instability. Under these ideal conditions, with high intrarater reliability of mean scores on each of the language aspects, all the criteria were met for convergent and discriminant valication of grammar, vocabulary, pronunciation, and fluency. Table 6 reveals no exceptions to the ideal requirements of convergent and discriminant validation of the four language aspects using mean scores on the TOP interview. Similarly, the correlated mean scores from Part C of the MLA speaking test presented in Table 7 show only one minor flaw: the correlation of the second rating of vocabulary with the second rating o grammar exceeds the correlation between first and second ratings of grammar by .001.

TABLE 5

Multitrait, Multimethod Convergent and Discriminant Validation Matrix

(N = 47 for all variables)

Test	Language Aspect	MLA Gr.	MLA Vo.	MLA Pr.	MLA Fl.	TOP Gr.	TOP Vo.	TOP Pr.	TOP Fl.
MLA	Grammar	----							
MLA	Vocabulary	.876	---						
MLA	Pronunciation	.882	.775	----					
MLA	Fluency	.845	.946	.731	----				
TOP	Grammar	.810	.827	.752	.783	----			
TOP	Vocabulary	.744	.816	.683	.796	.876	----		
TOP	Pronunciation	.741	.670	.788	.643	.838	.740	----	
TOP	Fluency	.687	.802	.657	.819	.864	.825	.731	----

Correlations in the validity diagonal are underlined.

All correlations in this matrix are significant at the p < .001 level.

TABLE 6

TOP Interview Multitrait, Multirating Convergent and Discriminant Validation Matrix

Test Rating	Language Aspect	1st Gr.	1st Vo.	1st Pr.	1st Fl.	2nd Gr.	2nd Vo.	2nd Pr.	2nd Fl.
First	Grammar	----							
First	Vocabulary	.876	----						
First	Pronunciation	.838	.740	----					
First	Fluency	.864	.825	.731	----				
Second	Grammar	.939	.832	.824	.829	----			
Second	Vocabulary	.883	.943	.799	.855	.891	----		
Second	Pronunciation	.829	.750	.909	.722	.810	.805	----	
Second	Fluency	.814	.716	.694	.908	.813	.791	.722	---

Correlations in the validity diagonal are underlined.

All correlations in this matrix are significant at the $p < .001$ level.

TABLE 7

MLA Speaking Test Multitrait, Multirating Convergent and Discriminant Validation Matrix

Test Rating	Language Aspect								
First	Grammar	----							
First	Vocabulary	.876	----						
First	Pronunciation	.882	.775	----					
First	Fluency	.845	.946	.731	----				
Second	Grammar	.937	.901	.837	.890	----			
Second	Vocabulary	.856	.953	.769	.915	.938	----		
Second	Pronunciation	.853	.758	.942	.743	.869	.802	----	
Second	Fluency	.795	.914	.707	.963	.886	.926	.739	----
		1st Gr.	1st Vo.	1st Pr.	1st Fl.	2nd Gr.	2nd Vo.	2nd Pr.	2nd Fl.

Correlations in the validity diagonal are underlined.

All correlations in this matrix are significant at the p < .001 level.

Conclusions

As shown in Table 8, ratings from TOP interviews were generally as reliable as scores on the MLA speaking test, indicating that an oral interview procedure can be developed that matches the reliability of the more structured MLA speaking test.

TABLE 8

Summary of Intraclass Reliability Coefficients for MLA and TOP Assessment Procedures

Test Score	Interrater Reliability	Intrarater Reliability	Test-retest Reliability
MLA Speaking Test Total Score	.818	.897	.940
MLA Part C Score	.829	.911	.940
Sums of Ratings from TOP Interviews	.827	.930	.893

It is also interesting that the reliability of Part C scores on the MLA test, which calls for free responses from examinees, was found to be as reliable as total MLA scores. Part C scores also correlated more highly with ratings from TOP interviews than did total MLA speaking scores. The product-moment correlation between Part C scores and sums of language aspect ratings from TOP interviews was .864, which approaches the test-retest reliability of the TOP interviews. Thus, Part C of the MLA test and the TOP interview seem to be generally measuring the same skill.

Interrater reliability was about equal for the MLA test and the TOP interview. Intrarater reliability was higher for the TOP interview than for the MLA speaking test, but for test-retest reliability the situation was reversed. This may have been the result of two factors. First, intrarater reliability may have been improved by the more detailed rating criteria used for rating the TOP interviews. Second, whereas the content of the MLA speaking test was exactly the same from one test administration to the next, TOP interviews were not identical in content. Adequacy of language content sampled may be a problem with both types of tests. The language sample provided by the MLA test is quite limited in scope, while the content of the TOP interview is dependent on the skill of the interviewer.

Correlations of ratings assigned the language aspects of grammar, vocabulary, pronunciation, and fluency using different testing and rating procedures ranged from .788 to .819. However, high correlations were found between different language aspects rated by the same method, which precluded convergent and discriminant validation of contributing language aspects across testing methods. This may indicate a halo effect among ratings assigned at the same time from the same speech sample, as well as variance resulting from different testing procedures and trait instability. Evidence of construct validity for the language aspects of grammar, vocabulary, pronunciation, and fluency was found by applying convergent and discriminant criteria to two independent ratings of the same test administration. Validity correlations consistently exceeded .90 for both testing procedures.

Implications and Recommendations for Further Study

The results of this study demonstrate that more direct measures of oral language proficiency may be as reliable as less direct but more structured standardized tests. The logical assumption that direct measures of oral language proficiency more accurately assess the skill being measured (Clark, 1972a) therefore indicates an advantage in testing by means of an interview. However, the high correlation of the MLA test results (especially Part C) with the interview ratings, combined with practical advantages in ease of administration offered by the MLA test, may make it an acceptable alternative in some situations.

Convergent and discriminant validation of grammar, vocabulary, pronunciation, and fluency ratings within testing procedures indicates that these aspects of oral language proficiency can be defined and measured reliably enough to provide a meaningful diagnostic profile of skills contributing to general oral proficiency.

Continued research should be conducted on the construct validity of the language aspects of grammar, vocabulary, pronunciation, and fluency to determine whether rating these language aspects independently with an intervening lapse of time may reduce the correlations found between different language aspects rated by the same method. Research should also be undertaken to determine if the language aspects of grammar and vocabulary may be more effectively tested with other assessment procedures, such as written tests.

References

Axelrod, Joseph. *The Education of the Modern Foreign Language Teacher for American Schools*. New York: The Modern Language Association, 1966.

Brière, Eugene J. "Current Trends in Second Language Testing." *TESOL Quarterly* 3 (1969):333-40.

_____________. "Are We Really Measuring Proficiency with Our Foreign Language Tests?" *Foreign Language Annals* 4 (1971):385-91.

Bryan, Miriam M. "MLA Foreign Language Profiency Tests for Teachers and Advanced Students." *The DFL Bulletin* 5,i (1965):4-7.

Buros, Oscar Krisen. *The Seventh Mental Measurements Yearbook*. Volume II. Highland Park, N.J.: Gryphon Press, 1972.

Campbell, Donald T., and Fiske, Donald W. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." In *Principles of Educational and Psychological Measurement*, edited by William A. Mehrens and Robert L. Ebel, pp. 273-302. Chicago: Rand McNally, 1967.

Carroll, John B. "Problems of Testing in Language Instruction: Some Principles of Language Testing." In *Report of the Fourth Annual Round Table Meeting on Linguistics and Language Teaching*, edited by Archibald A. Hill, pp. 6-10. Monograph Series on Languages and Linguistics. Washington, D.C.: Georgetown University Press, 1953.

_____________. "Foreign Language Proficiency Levels Attained by Language Majors near Graduation from College." *Foreign Language Annals* 1 (1967a):131-51.

_____________. *The Foreign Language Attainments of Language Majors in the Senior Year: A Survey Conducted in U.S. Colleges and Universities*. Cambridge, Mass.: Graduate School of Education, Harvard University, 1967b. [EDRS: ED 013 343.]

_____________. "The Psychology of Language Testing." In *Language Testing Symposium: A Psycholinguistic Approach*, edited by Alan Davies, pp. 46-69. London: Oxford University Press, 1968.

Clark, John L. D. *Foreign-Language Testing: Theory and Practice*. Philadelphia: The Center for Curriculum Development, 1972a.

_____________. "Measurement Implications of Recent Trends in Foreign Language Teaching." In *Foreign Language Education: A Reappraisal*, edited by Dale L. Lange and Charles J. James, pp. 219-57. The ACTFL Review of Foreign Language Education, Volume 4. Skokie, Ill.: National Textbook Company, 1972b.

______________. "Theoretical and Technical Considerations in Oral Proficiency Testing." In *Testing Language Proficiency*, edited by Randall L. Jones and Bernard Spolsky, pp. 10-28. Arlington, Va.: Center for Applied Linguistics, 1975.

Cooper, Robert L. "An Elaborated Language Testing Model." In *Problems in Foreign Language Testing*, edited by John A. Upshur and Julia Fata, pp. 57-72. Language Learning. Special Issue No. 3. Ann Arbor, Mich.: Research Club in Language Learning, 1968.

Harris, David P. *Testing English as a Second Language*. New York: McGraw-Hill, 1969.

Lado, Robert. *Language Testing: The Construction and Use of Foreign Language Tests*. London: Longmans, Green and Co. Ltd., 1961. Reprinted, New York: McGraw-Hill, 1965.

Lowe, Pardee, Jr. "Oral Proficiency Testing: How and Why?" Presented at University of Minnesota German Department Roundtable, February 13, 1975.

______________. "Oral Interview Applications, Problems and Research: A Survey." *Interview Testing Newsletter* 1 (1976):1-2.

Manual for Peace Corps Language Testers. Princeton, N.J.: Educational Testing Service, n.d.

MLA Foreign Language Proficiency Tests for Teachers and Advanced Students. Princeton, N.J.: Educational Testing Service, 1966. Now known as MLA Cooperative Foreign Language Proficiency Tests.

MLA Interpretation of Scores. Leaflet. Princeton, N.J.: Educational Testing Service, 1966.

Myers, Charles T., and Melton, Richard S. *A Study of the Relationship Between Scores on the MLA Foreign Language Proficiency Tests for Teachers and Advanced Students and Ratings of Teacher Competence*. Princeton, N.J.: Educational Testing Service, 1964. [EDRS: ED 011 750.]

"Qualifications for Secondary School Teachers of Modern Foreign Languages." *Publications of the Modern Language Association of America* 70, iv (1955):46-49.

Spolsky, Bernard. "Concluding Statement." In *Testing Language Proficiency*, edited by Randall L. Jones and Bernard Spolsky, pp. 139-43. Arlington, Va.: Center for Applied Linguistics, 1975a.

__________. "Language Testing--The Problem of Validation." In Papers on Language Testing 1967-1974, edited by Leslie Palmer and Bernard Spolsky, pp. 146-53. Washington, D.C.: Teachers of English to Speakers of Other Languages, 1975b.

Valette, Rebecca M. "Evaluation of Learning in a Second Language." In Handbook on Formative and Summative Evaluation of Student Learning, edited by Benjamin S. Bloom, J. Thomas Hastings, and George F. Madaus, pp. 815-53. New York: McGraw-Hill, 1971.

Wilds, Claudia P. "The Oral Interview Test." In Testing Language Proficiency, edited by Randall L. Jones and Bernard Spolsky, pp. 29-44. Arlington, Va.: Center for Applied Linguistics, 1975.

INTERVIEW TESTING RESEARCH AT EDUCATIONAL TESTING SERVICE

John L. D. Clark

Educational Testing Service

INTERVIEW TESTING RESEARCH AT EDUCATIONAL TESTING SERVICE

John L. D. Clark

Educational Testing Service has been involved in interview testing activities for about the past nine years. The first and largest of these activities is an ongoing project with the Peace Corps that began in 1969. During the first two years of the project, ETS language department staff--following an initial period of intensive training at the Foreign Service Institute--conducted a large number of interviews of Peace Corps trainees and volunteers, both in the U.S. and at in-country duty stations. For the past seven years, however, ETS collaboration with the Peace Corps has focused on the training of in-country Peace Corps personnel to conduct and rate interviews in the host country language, using an English-medium training program described in greater detail elsewhere in these proceedings.[1] To date, approximately 560 interview testers in 55 countries have been trained and certified under this program and have administered some 18,000 interviews.

A second program in which ETS has been participating involves the training of interview testers in English and French at the secondary school level in cooperation with the New Brunswick (Canada) Ministry of Education. This project is also described in greater detail, and from the perspective of a "front-line" New Brunswick interviewer, in a separate presentation.[2]

One recent project, while of a smaller overall scale than either the Peace Corps or the New Brunswick program, has permitted ETS to carry out a number of research studies and analyses in the areas of interview training, interview format, and scoring procedures that may be of interest to others using the interview technique or involved in the interpretation of interview results. This project derived from an interest on the part of the TOEFL (Test of English as a Foreign Language) program at ETS in the possibility of developing a test that could be used operationally within the TOEFL program as a measure of active speaking ability. Although the use of a direct, face-to-face interview would have been ideal from a theoretical standpoint, the cost and administrative complexity of offering this capability at each of the hundreds of TOEFL testing sites worldwide dictated the development of a tape-recorded test supplemented by a printed test booklet rather than a face-to-face test.

Even though a direct proficiency interview was not operationally feasible within the TOEFL program, the research committee overseeing the speaking test study recommended that a direct proficiency measure be used as the criterion instrument against which the less direct testing procedures could be compared and validated. It was further recommended

[1] See Lovelace paper, this volume.

[2] See Albert paper, this volume.

that, even before undertaking this portion of the study, the interview procedure itself be thoroughly investigated with respect to intra- and interrater reliability, the efficacy of the interviewer and rater training procedures, the effect of differing interview lengths, and related questions. These activities were carried out between January and March 1977 and produced the bulk of the experimental data reported here. Before presenting the study results, it will be useful to briefly describe the scope of the study and the specific procedures followed.

The basic procedural approach of the TOEFL study was to carry out, "from scratch," each of the activities involved in: the initial training of interviewers; interviewing under realistic administration conditions; and, finally, interview rating, both on-the-spot and at a later time by means of a tape recording made of each interview.

A total of four prospective interviewers were identified from a group of approximately twelve candidates, selection being made through inspection of resumes followed by personal interviews. All four interviewers were native speakers of English at the undergraduate or graduate level and had an excellent technical knowledge of English through various combinations of undergraduate and graduate level English study, graduate linguistics courses, and ESL teaching experience.

The training process for the four interviewers was essentially the same as for the Peace Corps and New Brunswick testers. Specifically, each interviewer attended an intensive two-day session in which ETS staff explained in detail the nature and operation of the interview and of the interview scoring procedure. Demonstration interviews were also conducted and critiqued as a group. During the late afternoon and evening of the two training days, each participant listened to a series of fifteen training tapes of interviews at score levels 0+ to 4+ to provide additional familiarization and practice with the scoring scale. The final step in the training process was to have each participant listen to and rate a second, randomized series of fifteen interviews for which the official score levels were not known in advance. For each trainee, the extent to which the trainee scores on all fifteen tapes corresponded with the official levels was taken as a measure of rating accuracy.

Approximately three and a half weeks after the initial training session, the four newly trained interviewers and the present writer carried out a three-day session of interview testing at the American Language Program (ALP) at Columbia University with a group of undergraduate and graduate students taking ESL courses at the ALP. A total of eighty-six students participated in the interviewing: forty-nine men and thirty-seven women, ranging in age from seventeen to sixty-one (S.D.=8.57) and representing twenty-six different languages.

The students were scheduled to appear for the interviewing over a three-day period at thirty-minute intervals. On arrival at the testing site, each student was asked to fill out a short questionnaire giving basic identification information. In addition, the student was asked to indicate his or her instructional level (present course placement) at the ALP and to give a self-rating of speaking proficiency on a 0-5 basis, using the regular verbal descriptions of each score level. This was accomplished by having the student read over each of the verbal descriptions and place a check mark opposite the description that was considered to best reflect his or her level of proficiency in spoken English. The questionnaire and self-rating information were put aside and were not seen by the interviewers at any point in the interviewing or rating process.

In order to explore the psychometric properties of an interview of appreciably shorter length than the usual (approximately twenty-minute) interview, each student was asked to participate in both a regular-length interview (hereafter, "long' interview) and a considerably abbreviated ("short") interview that was intended to run for a total of only five minutes. The order of interviewing was such that approximately half the students received the long interview first, followed immediately by the short interview, and half the short interview, followed immediately by the long. To avoid a carry-over or "halo" effect between long and short interviews, different interviewers were used to conduct the long and short interviews for a given student. Actual running times for the long interviews ranged from 10'10" to 26'27", with a mean duration of 18'6" and standard deviation of 3'43". The short interviews ranged in length from 4'20" to 8'54", with a mean of 6'33" and standard deviation of 1'8".

Over the three-day interviewing session, each interviewer conducted both long and short interviews for approximately equal total periods of time. Three interviewers began the session with long interviews and two with short interviews to counterbalance any sequence-of-interviewing effects across interviewers.

Both the long and short interviews were conducted on a one-interviewer-per-student basis, with no observers or "second raters" present. All interviews were cassette recorded, with small lapel microphones worn by the interviewer and the examinee. Immediately following the interview, the interviewer evaluated the examinee's performance, using the regular verbal criteria (including "pluses" where applicable) and noted this rating on the scoring form. However, the examinee was not informed of the rating at that time and the rating was not communicated in any way to the other interviewers.

The on-site interviewing sessions provided four basic types of examinee data:

1. the examinee's course placement at the ALP;
2. self-rating of speaking proficiency;
3. on-the-spot interview rating based on a long interview format;
4. on-the-spot interview rating based on a short interview format.

In addition to the above data, ALP staff made available each student's scores on a multiple-choice placement test administered by the ALP on entry to the language training p-ogram. This test consisted of a 60-item recorded listening comprehension section and a 120-item section covering English grammar and vocabulary. The placement test scores were not communicated to any of the interviewers until both on-site interviewing and rerating of the recorded interviews had been completed.

Approximately two weeks after the on-site interviewing session at ALP, each of the five interviewers listened to and rated all the tape recorded interviews, both long and short, including those he or she had given. The tapes were sequenced in such a way that, for each rater, approximately fifteen short interviews were followed by fifteen long interviews, or vice versa, until the rerating was completed. In no event were the long and short interviews for a given student listened to back-to-back; they were in all instances separated by at least fifteen intervening interviews. Discussions with the interviewers following the rerating process indicated that the raters could not remember individual examinees or the scores initially assigned, except for one or two examinees at the highest and lowest extremes of the score scale whose scores were remembered by the original rater because of the uniqueness of the performance. For all practical purposes, however, and because of the great number of interview tapes to be judged, the raters were not able to recollect the initially assigned scores when rerating the interviews.

On completion of the rerating phase, four further types of information were available to the study:

1. reratings of the regular long interviews by the original interviewer;
2. reratings of the short interviews by the original interviewer;
3. reratings of the long interviews by each of four additional raters;
4. reratings of the short interviews by each of four additional raters.

On the basis of the data obtained across the different phases of the study, it is possible to provide at least some empirically based information addressed to several different aspects of the interviewing and interview scoring process. To facilitate the presentation of results, generalized topical headings applicable to interview testing and research in a variety of contexts are used, followed by a description of study results bearing on that particular topic.

Tester Performance during Training and In-field Rating Accuracy

As previously described, each of the four interviewers trained for the TOEFL study was asked to rate a series of fifteen official test tapes, ranging from 0+ to 4+, as a measure of end-of-training rating accuracy. For each tape, the score given by the tester was compared to the official score. Trainee scores a "plus" above or below the official score (for example, an official 2 rated as a 2+ or a 4 rated as a 3+) received a discrepancy weight of plus or minus 0.5. Any scores given by a trainee that were one level above or below the official score received a discrepancy weight of plus or minus 1.0. For each tester, the discrepancies across all fifteen tapes were summed and both the absolute mean values and the signed mean values (taking into account the direction of the discrepancy as well as its magnitude) were determined, as shown in Table 1.

TABLE 1

Comparison of Rater Training Accuracy with Operational Scoring Accuracy

Rater	Absolute Mean Training Discrepancy*	Absolute Mean Deviation in Operational Rating
A	.40	.22
B	.50	.40
C	.10	.24
D	.30	.30

r = .603 (n.s.)

Rater	Signed Mean Training Discrepancy	Signed Mean Deviation in Operational Rating
A	.00	.10
B	-.27	.05
C	.10	.08
D	-1.00	-.01

r = .963 (p < .01)

*See text for definition of column entries.

As a measure of rating accuracy for each of the testers when working in an operational setting some weeks after training, the average of the ratings (across raters) given to each long interview during the relistening phase of the study was calculated. For each rater, the discrepancy of the rater's score from the average score for that interview was obtained. For each rater, the discrepancies across all interviews were summed and mean discrepancies, both absolute and signed, were calculated (right-hand column of Table 1).

A correlation of .603 was found between the absolute mean training discrepancy for a given rater and the corresponding absolute mean deviation in operational rating performance. With the small sample size (N=4), this correlation does not reach statistical significance. However, for the signed mean discrepancies, the obtained correlation was .963, significant at the p < .01 level and indicating a positive relationship between this end-of-training variable and interview scoring accuracy.

A caution in interpretation should, however, be noted. The elapsed time between initial training and operational scoring was relatively brief (approximately six weeks), and it is possible that the testers' scoring performance over a longer time period might exhibit variations from the initial training profile that were not in evidence over the period of the study. However, even taking this consideration into account, the obtained results for the signed discrepancy analysis would appear to provide a reasonable degree of validation for the use of this end-of-training measure as an indicator of probable rating performance in the field.

Intrarater Reliability

The TOEFL study provided some information on the intrarater reliability of the interview technique--that is, the extent to which individual raters "agree with themselves" when rescoring interviews to which they have earlier assigned ratings. Each of the five interviewers had initially interviewed approximately seventeen students face-to-face with the long interview format and approximately seventeen other students with the short format. During the rerating phase of the study, each interviewer listened to and rescored each of the interviews, long and short, that he or she had conducted, as well as those of the other interviewers. This activity provided intrarater reliability information for each of the raters, as shown in Table 2.

TABLE 2

Score-Rescore Reliabilities of Individual Raters

Long Interview			Short Interview		
Rater	r	N	Rater	r	N
A	.907	17	A	.837	17
B	.868	17	B	.904	14
C	.947	19	C	.853	15
D	.771	17	D	.740	18
E	.840	15	E	.751	11

For the long interviews, intrarater (or "score-rescore") reliabilities of .771 to .947 were obtained, with an average reliability of .867. Reliabilities for the short interviews were slightly lower, ranging from .740 to .904, with an average reliability of .817. In all but one instance, the short interview reliability for a given interviewer was slightly lower than the long interview reliability; the single exception was interviewer B, with long and short interview reliability figures of .868 and .904, respectively.

The intrarater reliability data also provide some information on the question of whether interview raters tend to evaluate examinee performance differently depending on whether the rating is carried out on-the-spot or is based on a tape recording of the interview that is listened to later. For both long and short interviews, the mean scores of each rater for both the initial (face-to-face) and subsequent (taped) ratings of those examinees he or she had interviewed are shown in Table 3. Nonsignificant differences in the mean scores for initial rating and rerating were found for raters A, B, and C in the long interview situation and for raters A and B in the short interview situation. However, for the long interviews, raters D and E assigned significantly higher scores (p < .05) to the rerated tapes than they had assigned during the face-to-face interviewing. For the short interviews, raters D and E were joined by rater C, who also gave significantly higher scores to the rerated tapes. Although the mean scores for the other rater/interview combinations did not vary significantly, in three of the four comparisons the numerical value of the mean was higher for the reratings.

TABLE 3

Mean Initial Ratings and Reratings
Assigned by Individual Testers

	Initial Rating			Rerating		
Rater	Mean	S.D.	N	Mean	S.D.	N
			Long Interview			
A	2.35	1.12	17	2.33	1.06	17
B	2.94	.75	17	3.02	.49	17
C	2.54	.80	19	2.68	.91	19
D	2.07*	.73	17	2.63*	.63	17
E	2.55*	.75	15	3.02*	.82	16
			Short Interview			
A	2.25	.81	17	2.34	.73	17
B	2.64	1.11	14	2.66	.75	14
C	2.41*	.59	15	2.69*	.84	15
D	2.41*	1.04	18	2.89*	.67	18
E	2.20*	.79	11	2.93*	1.04	11

*Initial and rerating means differ at p < .05.

From an operational standpoint, intrarater differences in scores assigned to tape-based ratings and on-the-spot ratings would not be a troublesome factor if the particular testing program utilized one of these two types of scoring procedures exclusively, that is, if reported scores involved only on-the-spot scoring or only tape-based scoring. However, for programs in which reported scores can include both on-the-spot and tape-based rating, it would appear desirable to carefully investigate possible rater differences due to the type of scoring procedure and to make allowance for any such differences in the use and interpretation of interview results.

Interrater Reliability

Interrater reliability refers to the extent to which two or more raters agree with one another on the scores they assign to given examinees. At ETS, data relating to the interrater reliability of the interview procedure have been obtained both from the TOEFL study and in connection with an interviewing program for Spanish-English bilingual and English-second-language teachers and teacher certification candidates in New Jersey.

In the TOEFL study, the five participating raters were asked to listen to and score a series of taped English interviews they and their four colleagues had conducted earlier on a face-to-face basis. A total of eighty-six long interview tapes were scored by all five raters. However, because of certain administrative problems in distributing the short interview recordings to the raters, it was not possible in several instances for all five raters to listen to and score a particular short interview. Interviews for which even a single rating was missing were removed from the analysis, leaving a total of sixty-eight short interviews for which complete scoring data (scores from all five raters) were available.

In the New Jersey study, four trained Spanish raters listened to and scored a total of eighty-six Spanish interviews drawn from the pool of interviews that had been conducted by the time of the study. For all three sets of data (long and short TOEFL English interviews and New Jersey Spanish interviews), intercorrelations of the scores assigned by the raters were calculated. These are shown in Table 4 together with the (arithmetic) mean correlation for each of the three correlation tables. As a general observation, it may be suggested that the obtained correlations for both the TOEFL and New Jersey data are within the overall levels of scoring reliability that would be expected for a nonobjective testing format of this type. The correlations also indicate that in all three scoring instances, the raters were able to rank the performance of the examinees whose interviews they evaluated in much the same way.

TABLE 4

Interrater Correlations for Three Sets of Recorded Interviews

TOEFL Interviews--Long (N=86)

Rater	A	B	C	D	E	
A	1.000					
B	.840	1.000				
C	.602	.705	1.000			
D	.780	.788	.712	1.000		
E	.814	.804	.593	.711	1.000	Mean r = .735

TOEFL Interviews--Short (N=68)

Rater	A	B	C	D	E	
A	1.000					
B	.857	1.000				
C	.778	.741	1.000			
D	.771	.767	.744	1.000		
E	.752	.782	.679	.709	1.000	Mean r = .758

New Jersey Interviews (N=86)

Rater	J	K	L	M	
J	1.000				
K	.900	1.000			
L	.775	.893	1.000		
M	.815	.854	.813	1.000	Mean r = .842

Although the correlation coefficients in Table 4 show a generally high correspondence of score rankings, they do not take into account possible absolute differences in assigned scores--that is, any tendency of individual raters to score a given examinee performance more leniently or more severely than their colleagues--even though they are in agreement on the relative rankings of the examinees. The question of possible differences in absolute scores was investigated by comparing the mean score ratings (across examinees) assigned by the raters in all three rating contexts; these results are shown in Table 5.

TABLE 5

Mean Interview Ratings for Individual Raters

TOEFL Interviews--Long (N=86)*

Rater	Mean Rating	S.D.
A	2.47	.82
E	2.67	.82
C	2.74	.89
D	2.77	.64
B	2.79	.70

TOEFL Interviews--Short (N=68)

A	2.41	.83
E	2.48	.85
C	2.54	.90
B	2.72	.63
D	2.76	.60

New Jersey Interviews (N=86)

L	3.70	.93
J	3.72	1.19
K	3.97	1.10
M	4.27	.80

*Raters sharing a common vertical line do not differ significantly in mean score (p > .05). Raters not joined by a line differ beyond p=.05.

For the long TOEFL interview ratings, the raters' mean scores ranged from 2.47 for the most severe rater to 2.79 for the most lenient. Ranges for the TOEFL short interview and for the New Jersey interview ratings were 2.41-2.76 and 3.70-4.27, respectively. The statistical significance of the difference in means between individual raters was determined through a series of t-tests for correlated means. The results of these tests are shown in Table 5 by means of vertical lines. Raters sharing a vertical line were not found to differ significantly in mean assigned ratings, while significant differences were obtained between raters not sharing a line.

Although these comparisons do show a number of statistically significant differences in the averages of the assigned ratings across raters, they do not of themselves provide a very useful or practical indication of the effect that scoring variability would be expected to have on the interview scores reported for individual examinees. This can be more readily determined by analyzing, for each examinee in a given scoring study, the interview ratings actually assigned by the raters and presenting this information in the form of expectancy tables showing the probability that an examinee whose reported score is at a given level would have a different scoring outcome if his or her performance had been evaluated by some other rater.

This approach is demonstrated in Table 6 for the New Jersey interview study. For each of three possible "passing score" levels shown in the table, observed frequencies and percentages of the same or different decisional outcomes are given. For example, if the passing score level is hypothetically set at 2+ (i.e., if all examinees scoring 2+ or higher are considered accepted and all those scoring below 2+ considered rejected), the middle of the three expectancy tables in Table 6 would be consulted. From these figures, based on the observed scoring performance of three additional raters beyond the initial rater, it can be seen that 82.6 percent of the additionally generated scores for examinees initially rated at level 2+ or higher were also 2+ or higher, and that 6.2 percent of the additional scores for examinees initially rated below level 2+ were also lower than 2+. By adding these two percentages (the upper left and lower right quadrants of the table), it may be seen that 88.8 percent of the reratings corroborated the initial decisional outcome as to acceptance or rejection at a level 2+ cutoff.

Percentages on the opposite diagonal indicate the proportion of rescorings in which the original outcome was not duplicated. Specifically, 11.2 percent of the reratings for interviews originally scored lower than 2+ were 2+ or higher, indicating that, in these instances, there was an 11.2 percent probability that the candidate would have had a favorable ("pass") outcome if he or she had been rated by another rater. Persons responsible for setting "passing" levels or making other kinds of decisions on the basis of the interview scores should take the nature and extent of scoring variability into account: in the example shown, consideration might be given to setting the passing score slightly lower than the initially intended level, to minimize the possibility that examinees who do in fact have the desired level of proficiency would be improperly rejected as a result of scoring variability of the interview process.

Relationship of Interview Scores to Other Indices of Language Competence

In addition to long and short interview scores for each examinee, available TOEFL project data included information on the instructional level of the English course to which the examinee had been assigned at the ALP, performance on the ALP placement test, and self-rating of speaking proficiency based on the regular interview scale.

TABLE 6

Expectancy Tables for Three Passing Score Levels (New Jersey Data)

Passing Score: 3 or Higher

A. Number of Scores

Reported Scores	Other Raters' Scores	
	3 or higher	lower than 3
3 or Higher	177	3
Lower than 3	45	33

B. Percent of Scores

Reported Scores	Other Raters' Scores	
	3 or higher	lower than 3
3 or Higher	68.6%	1.2%
Lower than 3	17.4%	12.8%

Percent Agreement = 81.4%

Passing Score: 2+ or Higher

A. Number of Scores

Reported Scores	Other Raters' Scores	
	2+ or higher	lower than 2+
2+ or Higher	213	0
Lower than 2+	29	16

B. Percent of Scores

Reported Scores	Other Raters' Scores	
	2+ or higher	lower than 2
2+ or Higher	82.6%	0.0%
Lower than 2+	11.2%	6.2%

Percent Agreement = 88.8%

Passing Score: 2 or Higher

A. Number of Scores

Reported Scores	Other Raters' Scores	
	2 or higher	lower than 2
2 or Higher	255	0
Lower than 2	0	3

B. Percent of Scores

Reported Scores	Other Raters' Scores	
	2 or higher	lower than 2
2 or Higher	98.8%	0.0%
Lower than 2	0.0%	1.2%

Percent Agreement = 100.0%

TABLE 7

Correlations of Long and Short
Interview Scores with
Other Indices of Language Competence

		(1)	(2)	(3)	(4)	(5)
1.	Instructional Level at ALP	1.000	.590	.558	.610	.551
2.	ALP Placement Test Score	.590	1.000	.348	.570	.707
3.	Self-Rating of Speaking Proficiency	.558	.348	1.000	.479	.430
4.	Long Interview Score	.610	.570	.479	1.000	.696
5.	Short Interview Score	.551	.707	.430	.696	1.000

The correlation matrix for all five of these variables is shown in Table 7. The lowest of these correlations (.348) is significantly different from zero (p < .01) and the highest correlations are well beyond .001. Although the greatest evidence for the validity of the interview technique as a measure of real-life speaking proficiency is considered to reside in the face and content validity of the procedure and the associated scoring scale, intercorrelations of the obtained interview scores with other kinds of language proficiency measures can provide some corroborating evidence.

With respect to the self-rating data, correlations of .479 and .430 for the long and short interviews, respectively, were found between the interview score results and student self-ratings of speaking ability using the regular FSI scale. Although these correlations are not extremely high, they suggest a clear positive relationship whose real magnitude is probably underrepresented to some extent as a function of measurement imprecision in both variables. Measurement precision of the student self-ratings could probably have been increased by allowing the students to indicate "plus" ratings where applicable, rather than rating on only the five broad numerical categories. In addition, simplification of and/or more detailed explanation of the meaning of each score category would probably have been helpful, especially for the less competent students, who may have encountered some difficulty in reading the verbal definitions of proficiency with full comprehension.

Although this approach is not possible in operational interviewing situations, a more precise estimate of the "true" interview scores for individual examinees in the TOEFL study may be obtained by averaging each of the five scores assigned by the interview raters when relistening to a given interview. Intercorrelations of long and short average interview scores with the self-ratings were found to be .560 and .554--an increase over the .479 and .430 correlations with the single interview rating, and presumably more indicative of the true extent of the relationship between the two variables after adjusting for the scoring unreliability of the interview.

Further experimentation with student self-ratings as related to obtained interview scores would provide extremely useful information about both the basic validity of the proficiency interviewing technique and the extent to which self-ratings of competence might in certain situations take the place of an externally administered interview. A major caution in this regard is that the examinee should be in a position to give a frank and honest appraisal of his or her level of proficiency. For situations in which it would be to the candidate's advantage to profess a higher (or lower) degree of competence than is actually the case, the self-rating technique would be of questionable validity and usefulness.

Another question of interest in the correlational data is the extent to which interview ratings might be used in place of typical multiple-choice testing procedures for instructional placement purposes. As shown in Table 7, the ALP placement test (consisting of 60 listening comprehension questions and 120 questions bearing on English grammar and vocabulary) correlated .590 with the instructional (class assignment) levels of the examinees at the time of the interviewing study. Correlations of .610 and .551 were found between the assigned instructional level and the long and short on-the-spot interview ratings. The three correlations do not differ significantly, indicating that both the long and the short interviews were able to predict assignment to instructional level as effectively as the multiple-choice placement test. Proponents of the interview technique might point out that even a quite abbreviated face-to-face interview lasting on the average only about six and a half minutes showed as much predictive power as the considerably longer and more time-consuming regular placement test. Proponents of more objective testing techniques might consider these results indicative of the extent to which testing procedures that do not require active speaking performance can substitute for direct measures in an operational placement context.

Length of Interview

The FSI-type interview is generally considered to require approximately twenty minutes of testing time for the majority of examinees and thirty minutes or more for examinees at the higher proficiency levels. Including the time required to greet the examinee at the beginning of the interview and to determine and record the interview rating following the interview, the overall testing time can be expected to work out to about thirty minutes per examinee, or no more than two examinees per hour. In light of the time and manpower requirements for interviews of the conventional length, there would be considerable practical value in reducing the total testing time per interview--provided this could be done without unduly affecting the face/content validity of the process or appreciably lowering the scoring reliability.

With respect to scoring reliability, data from the TOEFL study comparing both intrarater reliability (Table 2) and interrater reliability (Table 4) of regular length and considerably shorter interviews demonstrated little if any reduction in the reliability coefficients for the abbreviated interview format. As additional evidence, based on the mean interview rating across five raters, there was a correlation of .939 between the long and short interview scores for the TOEFL examinees, indicating a very high degree of underlying correspondence in the two variables. Further analyses are planned to determine the possible existence of interaction effects between score levels and scoring reliability--for example, the possibility that short interview scores are less reliably related to long interview scores at the upper end of the scoring scale than they are in the lower and middle ranges of the scale, where judgments based on a less extensive speech sample are presumably easier to make. Pending the detailed results of these analyses, the overall correlations obtained between long and short interviews would suggest that, at least from the standpoint of scoring reliability, interviews based on appreciably shorter running times merit serious practical attention.

With regard to the face/content validity of shorter-than-normal interviews (and including the psychological reactions of both interviewers and examinees to the reduced testing period), the TOEFL study interviews of approximately six and a half minutes average duration may be subject to discussion. Discounting the first half minute or so of both the long and short interviews, which is necessarily (and desirably) spent in greeting the examinee and exchanging a pleasantry or two, only about six minutes on the average were available under the short interview format for the interviewer to accomplish all the presumed necessary analytical tasks of the interview, that is, to establish the examinee's level of grammatical control, including tenses, agreements, and use of complex structures; extent of vocabulary as manifested in a variety of topical areas; and accuracy of pronunciation, overall fluency, and level of listening comprehension. Over the three-day interviewing period, many interviewers commented that, in the short interview situation, they would have liked to have had a bit more time with a number of the examinees and to have been able to ask a "few more questions" in order to make what they considered an adequate and confident judgment of the examinees' proficiency levels.

From the point of view of the examinee (in an other-than-experimental setting), an interview lasting no more than five to seven minutes might be viewed as inappropriately and unfairly short. Even though an accurate rating might indeed be possible in this length of time, the examinee could feel somewhat shortchanged in the conversational transaction and hence insufficiently probed as to overall proficiency.

An approach that would appear to maintain much of the practical and economic advantage of a short interview and at the same time provide for greater interviewer and examinee satisfaction in the length and scope of the procedure (as well as more fully support the face/content validity of the interview process) would be to make use of a medium-length interview of perhaps ten to twelve minutes, to be used with all but the most highly proficient examinees. Within this time period, and assuming that conversational digressions and overly long exploration of individual topical areas were kept to a minimum, the interviewer should be able to obtain a sufficiently extensive language sample to make an accurate rating and at the same time carry out a sufficiently wide-ranging conversation to satisfy the affective expectations of the process.

If procedures could be developed to carry out the entire interviewing and rating sequence for a majority of examinees within a fifteen-minute rather than a thirty-minute period, the total testing time for large numbers of examinees would be effectively halved, with concomitant savings in manpower and testing costs. For situations in which total testing time is not a significant concern (as, for example, in relatively low-volume testing carried out on an as-needed basis by regular members of an institutional staff), twenty-minute or longer interviews could of course be utilized and justified on both measurement and economic grounds. In other situations involving large numbers of examinees, outside interviewers, or other significant time/cost factors, a shorter interview format optimizing both validity/reliability and manpower/cost factors would merit serious consideration. Present indications from available ETS data are that a considerable reduction in total interviewing time should be possible without adversely affecting the scoring reliability or linguistic integrity of the process.

PSYCHOPHYSICAL SCALING OF THE LANGUAGE PROFICIENCY INTERVIEW

A PRELIMINARY REPORT

Robert J. Vincent

Central Intelligence Agency

PSYCHOPHYSICAL SCALING OF THE LANGUAGE PROFICIENCY INTERVIEW[1]

Robert J. Vincent

Background

Few language teachers or researchers would be expected to argue with the statement that for a given foreign language, a beginning student would experience more difficulty achieving a 3+ level on the eleven-point Foreign Service Institute (FSI) speaking proficiency scale than he would in reaching, say, the 2 level. But would the same teachers or researchers agree if asked to judge how much more difficult the 3+ level is to achieve than is the 2 level?

How much consensus would there be to a more complicated set of questions? Which is more difficult to achieve, and how much more: reaching a 3 level from a 0+, or a 4 from a 3+? How long should the average student in each category be enrolled in training? Is it possible to project from known durations of training to situations where, as yet, no data exist?

These and similar kinds of questions have cropped up time and again during a series of joint research efforts by the Psychological Services Staff (PSS) and the Language School (LS) to predict the speaking efficiency of language students at the conclusion of training. To be perfectly candid, we are rather proud of our ability to prognosticate on the basis of selected linguistic and psychological variables. Yet one thing we have learned along the way: the only two variables common to all of the languages investigated thus far are duration of training and speaking proficiency at the outset of training.

These recurring findings, coupled with the thought provocations just advanced, have led to a search for a unitized measure or scale of the difficulty of learning a foreign language.

[1] The author wishes to recognize the unusual measure of support given this research by the judges and Language School management personnel. They gave generously of their time, patience, enthusiasm, and expertise (despite certain misgivings about the stability of the author for having them throw numbers about in such an unorthodox fashion). A special word of thanks is due Dr. Pardee Lowe, Chief of Testing of the Language School. Without his complete cooperation--which by now the author very much takes for granted--neither this nor any of the other joint research projects conducted over the past several years would have come to fruition.

The need for scales traces its ancestry to the laboratories of German and French physicists. While it is true that the ancient Greeks sought laws relating the responses of man to the world around him, the Europeans made the first significant breakthroughs in relating sensory attributes such as loudness and brightness to their corresponding physical attributes: dynes per cm^2 and lamberts. These endeavors evolved into a branch of psychology referred to as psychophysics. Stevens (1936), considered by many to be the father of modern psychophysics, embarked on a vigorous, forty-year program to scale a variety of sensory continua. A prolific and at times irascible spokesman, his initial efforts were generated by a commercial requirement for a scale of subjective loudness. The physical (decibel) scale did not behave at all like its psychophysical (loudness) counterpart--simply put, 50 db does not sound half as loud as 100 db. Hence, the communication engineer needed a scale whose numbers made more sense to his customers than did the numbers on the decibel scale. The result was the sone scale (Stevens, 1955) which was subsequently adopted by the International Standards Organization to describe loudness for engineering purposes.

Psychophysicists were content to occupy themselves with true sensory problems until the mid 1950s. By that time they had reached general (but by no means universal) agreement on a psychophysical law: for nearly three dozen sense modalities (such as loudness, brightness, taste, heaviness, judged intensity of electric shock, and so forth), equal stimulus ratios produce equal perceptual ratios. Expressed mathematically:

$$\Psi = k(\Phi - \Phi_o)^n$$

where the perceived magnitude Ψ grows as the physical scale Φ raised to a power _n_. The Φ_o is often thought of as a threshold, while _k_ is merely a constant that depends upon the units employed. One particularly useful feature of this law is that when log Ψ is plotted against log Φ, the resulting power function is a straight line. Most importantly, each of the modalities abiding by the law seems to have a characteristic exponent (_n_), ranging from 0.3 for brightness to 3.5 for apparent intensity of electric shock (Stevens, 1961).

In the late 1950s the psychophysical techniques that had been found to work so well on measurable, physical (metric) continua began to be applied to stimuli that could be described only on a nominal (nonmetric) scale--attitudes, verbal statements, occupations, crimes, punishment, and musical selections, to name just a few (Stevens, 1966). Interestingly enough, the psychophysical power law seems to have held. Without some sort of metric, of course, the law could not be directly confirmed, but in the several instances where corresponding metrics were subsequently scaled, the relationship between judgments and physics entailed a power law.

Given this background, it seemed worthwhile to bring the psychophysical tools to bear on the matter of scaling the difficulty of learning a foreign language. This paper summarizes the extent to which this goal has been achieved.

Method

Eighteen faculty members of the LS volunteered to participate in the research. Each was asked to judge the difficulty the "average" LS student experiences in achieving the various speaking proficiency levels of the eleven-point FSI scale. The specific methods by which they went about this task are discussed in the next section. Suffice it to say at this point that judgments were restricted to the single foreign language the rater considered to be his area of prime expertise. The language categories included French, Spanish, German, Russian, Chinese (Mandarin), Japanese, Swedish, Arabic, Turkish, Portuguese (Brazilian), and Indonesian. Results from four participants were excluded from the analysis because the judges did not fully comply with the instructions, or because they were unable to complete the task due to prior commitments.

Two methods for judging the difficulty of learning to speak foreign languages were employed in the study. Copies of the instructions and response forms may be found in Appendix B.

Phase 1--Magnitude Estimation. The most direct and perhaps most efficient method to obtain an estimate of the relation between the FSI scale and judged difficulty attendant with reaching a particular FSI level is by means of magnitude estimation. The technique was employed as follows: a list of all eleven FSI levels was presented to each judge. Heading the list was a 2+ (the midpoint of the FSI scale), which was referred to as the "standard." An arbitrary number of 10 was assigned to it to describe its relative difficulty to achieve at the conclusion of training. Each of the remaining ten comparison FSI levels (arrayed in a different randomized order for each participant) was then judged by having the participants decide what number should be assigned to describe its difficulty to achieve relative to the 2+ standard. For example, if a particular FSI level was judged to be three times more difficult than a 2+, it received a value of 30. If another level was considered only one-tenth as difficult, it was called a 1, and so on.

The method of magnitude estimation was deliberately chosen as the lead-off technique because it is relatively straightforward and usually easily understood. Language School administrators had cautioned that some participants could be expected to experience difficulty interpreting the instructions because English was not their native language. As it turned out, few participants voiced any concern whatsoever, and nearly all completed Phase 1 in the allotted time of fifteen minutes. Several judges did express reservations, noting that they disliked working with numbers and that their results would be meaningless (a typical reaction in this kind of research). Nonetheless, they were encouraged to try and, with few exceptions, produced results entirely in keeping with those of the remaining judges.

Phase 2--Ratio Estimation. A second psychophysical technique was employed for several reasons. In the first place, despite the preliminary nature of this research, some means for independent verification of the results seemed to be in order. Second, the magnitude estimation method was limited by virtue of the fact that, as it was employed in this study, it focused on the "average" student's exit proficiency (that is, his FSI rating at the conclusion of training). Since it did not directly account for the fact that students can enter training at any FSI level (enter proficiency), the judges were left with the following options: either restrict their judgments to the case where enter proficiency was assumed to be 0, or somehow mentally average across all possible enter proficiencies to arrive at a single number appropriate to the exit proficiency in question.

The method of ratio estimation solved both problems. If indeed judged difficulty obeys the power law, both psychophysical techniques should produce similar results, with one serving as a check on the other. Moreover, the ratio estimation technique required the judges to assign numbers to all possible combinations of pairs of enter and exit proficiencies (excluding those cases where the enter proficiency scores equaled or exceeded exit proficiency scores). An enter score of 1+ and an exit score of 3 were chosen to represent the standard of 10. All remaining, randomized pairings were then judged relative to the standard pair. The judges were simply instructed to assign to the comparison pairs numbers proportional to the relative difficulty of the standard pair. Whereas magnitude estimation involved only exit proficiencies, ratio estimation was concerned with pairs of proficiencies. Otherwise, the scaling techniques were similar.

For the record, the judges found the ratio estimations much more of a challenge, and several took the opportunity to say so in no uncertain terms. If their magnitude estimates were meaningless, they noted, their ratio estimates had to be worse. As before, the experimenter attempted to assuage their concerns and asked them to do their best. Although most judges completed the task in the allotted forty-five minutes, some required twice as much time.

Results and Discussion

The experiment was expressly designed so as not to constrain the participants' definition of what constituted difficulty of learning to speak a foreign language. As a case in point, no mention was ever made by the experimenter that one way to assess the relative difficulty of the various FSI proficiency levels would be to compare the average durations of training associated with each combination of enter and exit proficiency ratings. Indeed, both the formal instructions as well as the informal introductory remarks stressed that difficulty was a judgment and that its definition probably varied from person to person and language to language. The experimenter expressed sympathy with how strange it must seem to be asked to assign numbers to such a nebulous dimension. Interestingly enough, not one participant volunteered that estimated duration of training constituted the basis for his judgments of difficulty (although that in no way discounts the possibility that duration was, in fact, the basis).

In any event, when the judges' estimates of difficulty of achieving each FSI level were compared to the average duration of training required to achieve that level, the resulting functions offered surprisingly strong confirmation of the psychophysical power law (Figure 1).[2] As a matter of fact, judged difficulty was described by both psychophysical methods as being directly proportional to the duration of training.[3] In mathematical terms,

$$\Psi = k(\Phi - \Phi_0)^n$$

where Ψ refers to estimated difficulty, $\underline{k}$ is a constant with a value of .01 or .03, depending on the psychophysical technique, Φ is duration of training, Φ_0 is a constant with a value of 37 for magnitude estimation and 0 for ratio estimation, and $\underline{n}$ = 1.00 for magnitude estimation and 1.03 for ratio estimation.

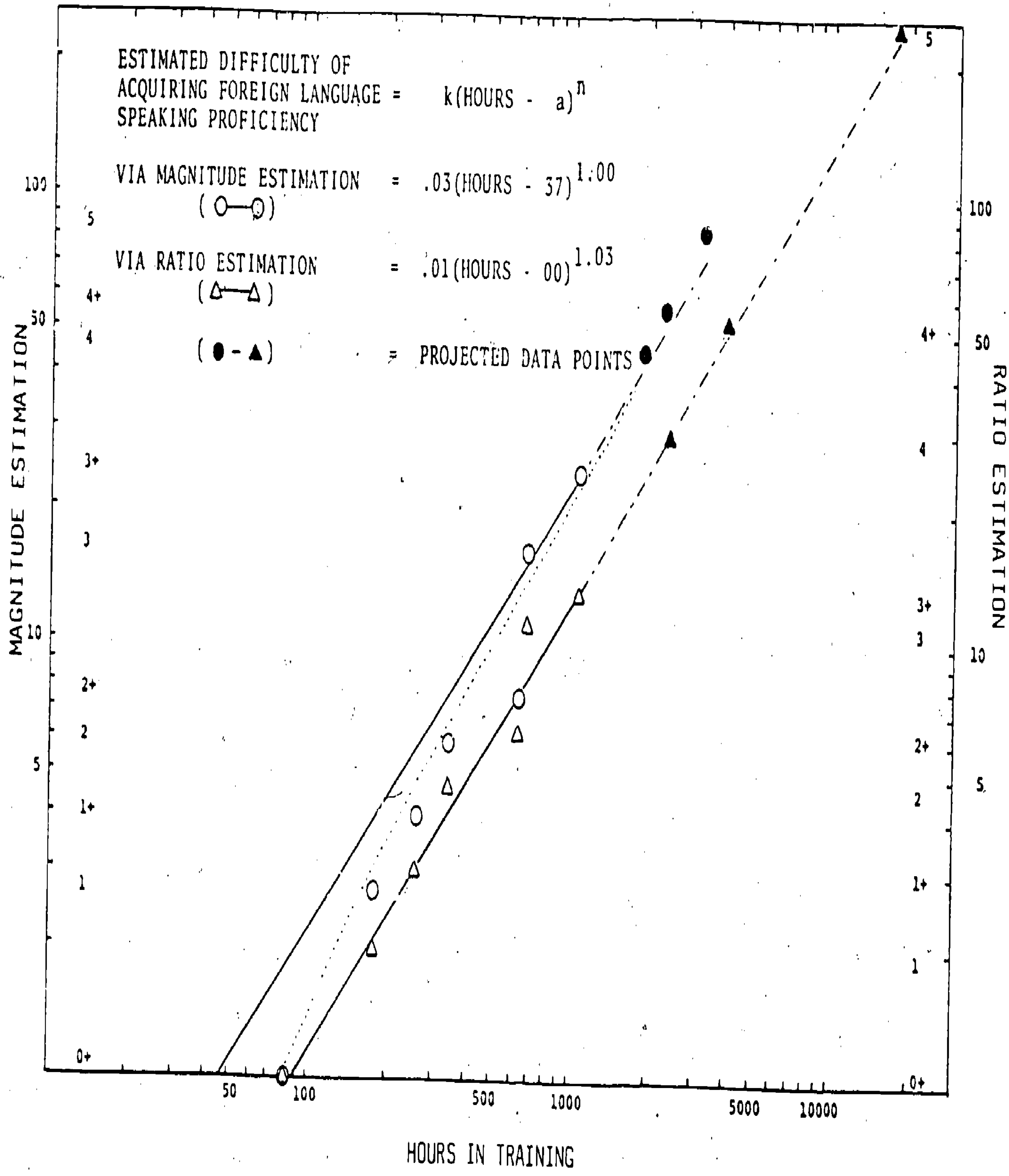
Adhering to standard procedures for handling highly variable data of the type found in psychophysical studies (Stevens, 1960), geometric means rather than arithmetic means were calculated for each enter and exit proficiency combination. This was true for both the judges' estimates of difficulty and the empirical durations of training.

Table 1 summarizes the duration of training data for the six languages in the data base. It should be mentioned that for the higher enter/exit combinations, few data points were available for use, and inspection of Table 1 reveals that no data whatsoever existed for the categories beyond 3+. Security considerations prevent disclosure of the numbers of students or measures of the variability of the data falling within each category.

Statistical procedures formulated by Ekman (1961), Mashhour (1961), and Torgerson (1958) were followed in deriving the two psychophysical scales. The power functions were calculated exclusively on the basis of training duration data found in the 0 through 3+ FSI categories; training durations associated with the 4, 4+, and 5 levels were then projected on the basis of the resulting power functions (and shown as filled data points in Figure 1).

[2] Duration of training data were compiled from the PSS computerized data base for LS students enrolled since 1969 in French, Spanish, German, Russian, Chinese, and Japanese.

[3] See Note 1, Appendix A.



ESTIMATED DIFFICULTY OF ACQUIRING FOREIGN LANGUAGE SPEAKING PROFICIENCY = k(HOURS - a)^n
VIA MAGNITUDE ESTIMATION = .03(HOURS - 37)^1.00
(O—O)
VIA RATIO ESTIMATION = .01(HOURS - 00)^1.03
(Δ—Δ)
(● - ▲) = PROJECTED DATA POINTS
MAGNITUDE ESTIMATION
100
50
10
5
5
4+
4
3+
3
2+
2
1+
1
0+
RATIO ESTIMATION
100
50
10
5
5
4+
4
3+
3
2+
2
1+
1
0+
50
100
500
1000
5000
10000
HOURS IN TRAINING

FIGURE 1

TABLE 1

Consecutive Weeks in Language Training*
(Empirical Data)

ENTER PROFICIENCY

EXIT PROFICIENCY	0	0+	1	1+	2	2+	3	3+	4	4+
0+	2.4									
1	6.7	2.7								
1+	12.1	4.5	4.1							
2	13.9	10.5	7.2	4.4						
2+	17.7	26.2	9.4	6.1	3.9					
3	18.5	16.6	9.9	9.7	9.4	4.5				
3+	29.0	11.5	--	--	16.5	4.2	3.1			
4	--	--	--	--	--	--	--	--		
4+	--	--	--	--	--	--	--	--	--	
5	--	--	--	--	--	--	--	--	--	--

*Based upon data available on French, Spanish, German, Russian, Chinese (Mandarin), and Japanese training programs.

The specific difficulty scale derived from magnitude estimations was found to be:

$$\text{Estimated Difficulty} = .03\,(\text{Hours in training} - 37\text{ hours})^{1.00}.$$

The comparable function for the ratio estimations was:

$$\text{Estimated Difficulty} = .01\,(\text{Hours in training})^{1.03}.$$

Recall from an earlier discussion that the .03 and .01 values are simply constants that move the functions up and down the scale according to the units of measurement chosen by the judges. Beyond that, they are of little interest to the discussion at hand.

The thirty-seven hour figure in the magnitude estimation function is another constant, and is often thought of as a noise threshold in the pure psychophysical studies (although even there its lineage occasionally is indeterminate). Mathematically, it serves to straighten out an otherwise curvilinear function. Whereas no such constant was required for the ratio estimation data, inspection of Figure 1 reveals that the magnitude estimation function would have been markedly curvilinear had not the constant been taken into account. For present purposes this additive constant will be viewed as a statistical expedient for curve fitting purposes, since the areas of prime interest rest with the overall relationship of judged difficulty to duration of training, and especially the slopes of these linear relationships.

But we would be remiss not to point out (at least parenthetically) that the thirty-seven hour constant is nearly identical to the average number of hours spent in training by those LS students who entered at, but were unable to progress beyond, the 0 proficiency level.

Note also in Figure 1 that the corresponding FSI ratings have been plotted along the estimation axes. These results are interpreted as follows: according to the magnitude estimation scale, an FSI level of 5 was judged to be about 85 times more difficult to achieve than a 0+, but only twice as difficult as a 4. Looking over to the ratio estimations, a 5 was estimated to be about 240 times more difficult to reach than a 0+, and more than 8 times more difficult than a 4. In other words, although the overall relationship between judged difficulty and duration of training obtained by two procedures was described by nearly identical power functions, the respective ranges of difficulty and the distribution of FSI levels within each range differed according to the psychophysical technique chosen. The differences between the two techniques are most striking at the 4 and higher levels. The magnitude estimation scale suggests that a student can achieve a 4 rating in approximately 3,250 hours (about 88 weeks), whereas the ratio estimation scale projects nearly 18,000 hours (or nearly 9.5 years). It is doubtful that many instructors would be as optimistic as the magnitude estimation projection, and the ratio estimation projection may be too low as well. But at least it squares with the opinion of some linguists that languge proficiency is fairly well established by the age of ten (Chomsky, 1968).

In any event, results discussed thus far appear to have satisfied two of the goals set forth for this research: scaling difficulty of learning a foreign language, and relating this difficulty to duration of training.

Carrying the analysis a step further, it was possible to use the power fun· ions to project the average number of hours in training for every combination of enter and exit proficiency. Two such projections have been made. The left-hand and center scales of Figure 2 show once again the relationship of FSI levels to estimated difficulty. These results came from the Phase 2 (ratio estimation) study and are identical to those depicted on the right side of Figure 1. Magnitude estimation data could have been used as well, but they were not, owing to the abbreviated range of judgments and the fact that, as mentioned earlier, such estimates were based upon overall estimates of the difficulty of exit proficiency rather than upon pairs of enter and exit proficiencies.

The right-hand scale is an artificial difficulty scale specifically calculated to even out the differences found among the various FSI levels on the ratio estimation scale. For example, the original scale (left side) indicates that the difference between a 4+ and a 5 is considerably larger than, say, the difference between a 4 and a 4+, despite the fact that the results are already plotted on a logarithmic scale (which would guarantee that even if the three levels had fallen equidistantly from one another, the relative difficulties would increase logarithmically). Four possibilities can be thought of as accounting for these disproportionalities: (1) the difficulty estimates are accurate--a 4+ is in reality very much less difficult to reach than is a 5, but only moderately more difficult than is a 4; (2) the judges had trouble estimating the difficulty of the levels, especially the mid- and upper-range levels; (3) the variations among levels could reflect how appropriately the judges regarded and were able to use numbers and ratios; or (4) some combination of these factors was at work.

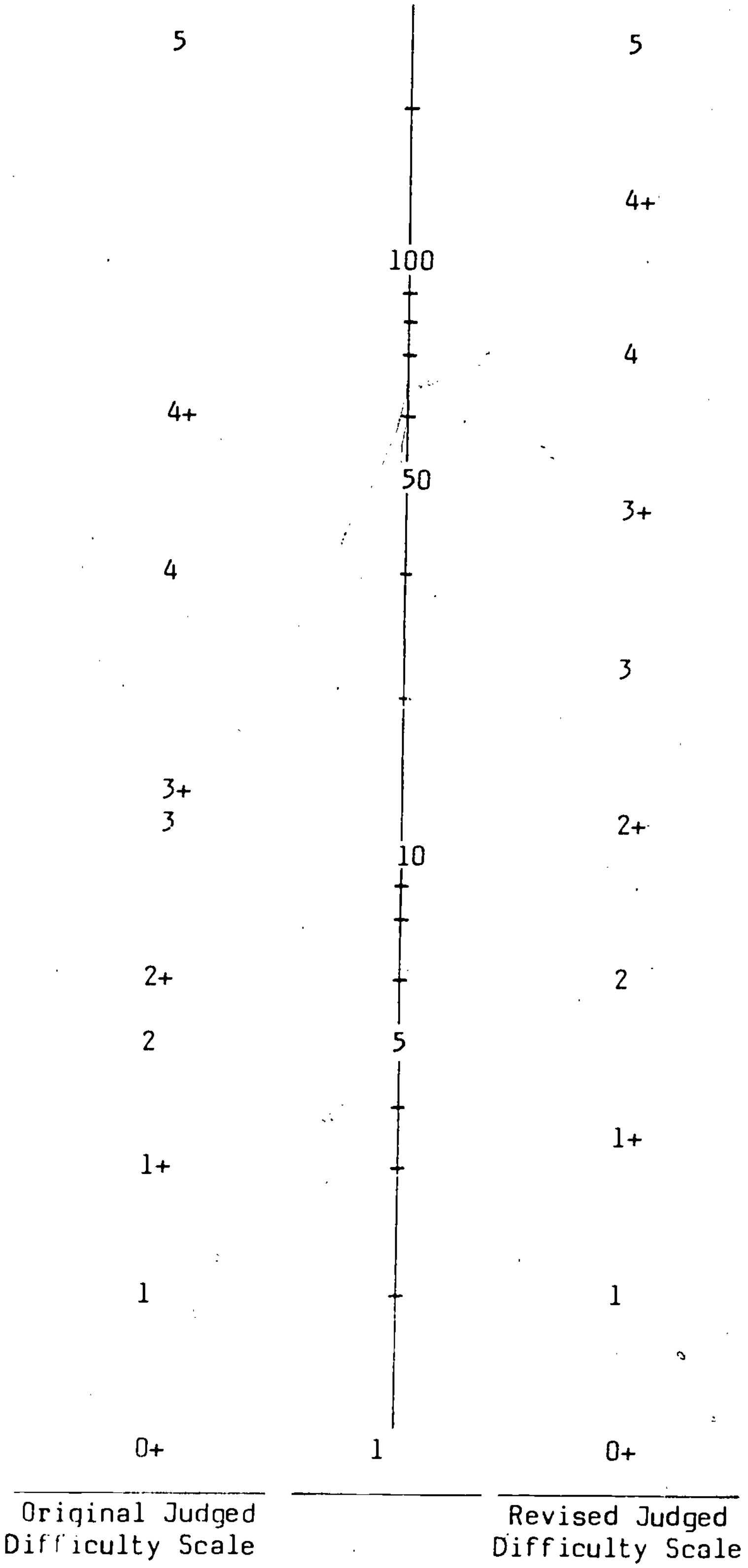
While the last (or compromise) hypothesis probably covers all the bases, the second hypothesis more than likely focuses on the single most significant contributor to respose variability. Nearly all judges remarked that they had never trained an adult student beyond the 4 or 4+ level (in some cases, beyond a 3+), and therefore could not imagine how difficult a task it would be, assuming that it were at all possible.

Although there is no a priori basis for accepting either scale as it applies to the higher FSI levels (recall that no data existed in our computerized records for the 4 through 5 levels) each scale could be compared to the empirical data base in the 0 through 3+ levels (Table 1). Such a comparison presumes acceptance of the data base as representing learning to speak a foreign language in general, despite the fact that some data points were based upon very small numbers of students (who themselves may or may not have been representative of students in general). In addition, all data points reflect most heavily the influence of students of French and Spanish, less heavily German, Russian, Chinese, and Japanese, but no other languages. About all that can be said in defense of the data base is that it represented the totality of information on duration of training available at the time this study was

FIGURE 2

Original and Revised Judged Difficulty Scales

(see text)



conducted. To the extent that it does adequately reflect how long the average student spends in training, it can be expected to provide useful results.

To this end, projected durations of training for each combination of enter and exit speaking proficiencies were compiled according to the ratio estimation power function.[3] The results are displayed in Table 2, with the original scale results posted at the top and the revised (log equidistant scale) results at the bottom. A comparison of these results with the empirical data in Table 1 is summarized in Figure 3. With a few rather conspicuous exceptions (such as 0+ to 2+ and 0+ to 3+), the judges' original estimates were reasonably accurate reflections of the actual durations of training in each enter/exit category. On the average, the original scale overestimated duration of training up to the 3+ level by approximately 1.3 weeks, whereas the revised scale overestimated training duration by more than 17 weeks. In short, the results support the contention that FSI levels are not spaced equidistantly along a logarithmic scale. Some levels are very much more or less difficult to achieve than would be predicted by a linear or logarithmic projection.

Finally, in answer to the question posed earlier (Which is more difficult to achieve, and how much more: reaching a 3 level from a 0+, or a 4 from a 3+?), note once again the top portion of Table 2. The projected duration for the former case is twenty-two weeks, compared to thirty-four weeks for the latter. Thus, progressing from a 3+ to a 4 is projected to take 1.5 times longer than advancing from a 0+ to a 3.[4] The empirical training data (Table 1) led to a dead end, since no 4+ data are cited. However, some last-minute detective work uncovered the records of several students who satisfied the 3+/4 requirement. Their average duration of training was, surprisingly, only 18.5 weeks, resulting in a 1.1 to 1 ratio for the empirical data. In either case, 3+ to 4 shows every indication of being more difficult to achieve than a 0+ to 3.

[3] See Note 2, Appendix A.

[4] See Note 3, Appendix A.

TABLE 2

Projected Consecutive Weeks in Language Training*

(Based Upon Judged Difficulty = .01 [Hours] 1.03)**

	ENTER PROFICIENCY									
EXIT PROFICIENCY	0	0+	1	1+	2	2+	3	3+	4	4+
0+	2.4									
1	4.6	2.2								
1+	7.0	4.6	2.4							
2	10.8	8.4	6.2	3.8						
2+	13.9	11.6	9.4	7.0	3.2					
3	24.4	22.0	19.8	17.4	13.6	10.4				
3+	28.8	26.5	24.2	21.9	18.1	14.9	4.5			
4	62.9	60.5	58.3	55.9	52.1	48.9	38.5	34.0		
4+	111.4	109.0	106.8	104.4	100.6	97.4	87.0	82.5	48.5	
5	484.6	482.2	480.0	477.6	473.6	470.6	460.2	455.7	421.7	373.2

A. Based upon Original Judged Difficulty Scale

	ENTER PROFICIENCY									
EXIT PROFICIENCY	0	0+	1	1+	2	2+	3	3+	4	4+
0+	2.4									
1	4.3	1.9								
1+	7.7	5.4	3.4							
2	13.9	11.6	9.7	6.2						
2+	25.2	22.8	20.9	17.5	11.2					
3	45.5	43.1	41.2	37.8	31.6	20.3				
3+	82.2	79.8	77.9	74.5	68.3	57.0	36.7			
4	148.5	146.2	144.3	140.8	134.6	123.3	103.0	66.3		
4+	268.3	266.0	264.1	260.6	254.4	243.2	222.8	186.1	119.8	
5	484.8	482.5	480.6	477.1	470.9	459.6	439.3	402.6	336.3	216.5

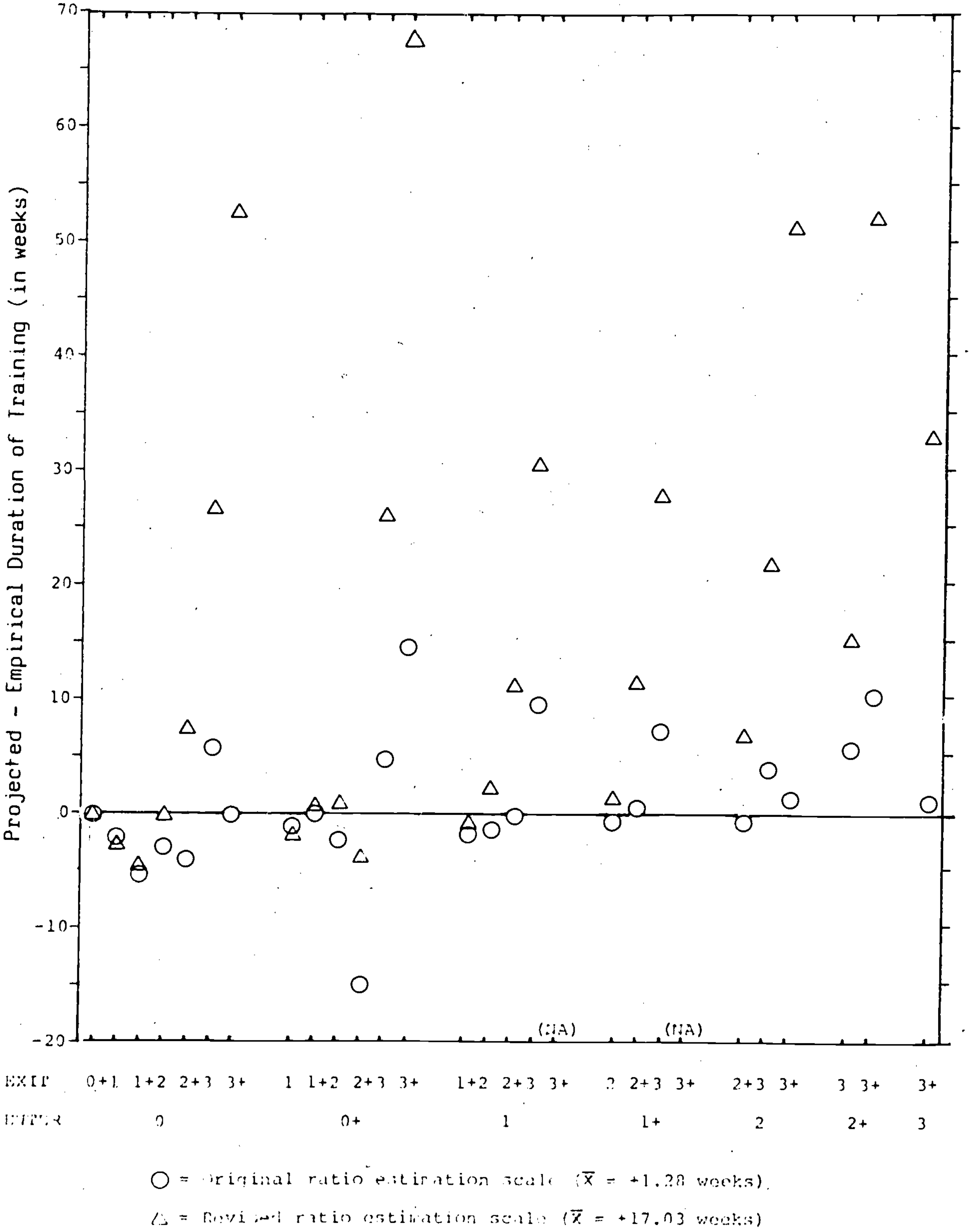
B. Based Upon Revised Judged Difficulty Scale.

*Based upon estimates of difficulty by instructors in French, Spanish, German, Russian, Chinese (Mandarin), Japanese, Portuguese (Brazilian), Swedish, Turkish, Arabic, Indonesian.

**Duration estimates based upon data available on French, Spanish, German, Russian, Chinese (Mandarin), and Japanese training programs.

FIGURE 3
Deviation of Projected from Empirical Duration of Training

(for enter/exit pairings, 0 through 3+)



Conclusions and Recommendations

Preliminary though they may be, the results rather strongly suggest that judged difficulty in learning to speak a foreign language can be scaled, and that difficulty is directly related to duration of training by the psychophysical power law. Since the law permits one to state that equal stimulus ratios produce equal perceptual ratios, it was possible to apply the judges' estimates to projections beyond available data, thereby generating a complete matrix of duration estimates for all pairs of enter and exit speaking proficiencies. It was further concluded that the estimated difficulty of achieving sequential FSI levels is not a straightforward progression. Some levels, especially those beyond the 2 or 2+ level, seem to require disproportionate amounts of training.

In recognition of the preliminary nature of this study, it is recommended that further work be pursued, with particular emphasis on enlarging the data base to include a wider selection of languages; filling in the gaps in the empirical duration-of-training data base; determining if individual languages obey the power law and, if so, grouping them according to their relative judged difficulty (and comparing the resulting groupings with those currently available); and, finally, calculating the judged difficulty of learning to read and understand foreign languages.

Appendix A

NOTE 1

Psychophysical Power Law

$$\Psi = k(\Phi - \Phi_o)^n$$

where:

Ψ = Perceived Magnitude (Judged Difficulty)

k = Constant

Φ = Physical Magnitude (Duration of Training)

Φ_o = "Threshold" (Statistical Expedient)

n = Exponent (Unique to Given Modality)

NOTE 2

Ratio Estimation

Example:

S.ENTER = 2+, S.EXIT = 4

$$\frac{\Psi_4}{\Psi_{2+}} = \left(\frac{\Phi_{2+ \rightarrow 4} + \Phi_{0 \rightarrow 2+}}{\Phi_{0 \rightarrow 2+}} \right)^n$$

where:

Ψ_4 = Judged Difficulty: S.EXIT = 4

Ψ_{2+} = Judged Difficulty: S.ENTER = 2+

$\Phi_{2+ \rightarrow 4}$ = Duration: S.ENTER = 2+; S.EXIT = 4

$\Phi_{0 \rightarrow 2+}$ = Duration: S.ENTER = 0; S.EXIT = 2+

n = 1.03

NOTE 3

Relative Difficulty

Compare:

S.ENTER = 0+; S.EXIT = 3

S.ENTER = 3+; S.EXIT = 4

$$\frac{\psi_{0+\to 3}}{\psi_{3+\to 4}} = \left(\frac{\phi_{0+\to 3}}{\phi_{3+\to 4}}\right)^{n}$$

where:

$\psi_{0+\to 3}$ = Judged Difficulty: S.ENTER = 0+; S.EXIT = 3

$\psi_{3+\to 4}$ = Judged Difficulty: S.ENTER = 3+; S.EXIT = 4

$\phi_{0+\to 3}$ = Duration: S.ENTER = 0+; S.EXIT = 3

$\phi_{3+\to 4}$ = Duration: S.ENTER = 3+; S.EXIT = 4

n = 1.03

Appendix B

Instructions, Phase 1

On the next page is a list of speaking exit proficiency ratings. Your task is to judge the difficulty you would expect the average LS student to experience in achieving each rating. You are to express this difficulty by assigning numbers to the ratings. The first rating, a 2+, is to be called "10." Thereafter, you are to assign numbers proportional to your subjective impression of this first rating. For example, if you feel a particular exit rating is twice as difficult to achieve as a 2+, assign to it a number "20." If you judge another to be one-fifth as difficult, call it "2," and so forth. Please do not restrict your response range. Use numbers as large or as small as you feel are necessary, including those less than "1" (fractions or decimals) if they are appropriate. Base your judgments on a specific foreign language with which you have had extensive teaching experience. Please note at the bottom of the list which language you had in mind.

SPEAKING
EXIT

2+ 10

2 ____

0 ____

1+ ____

2+ ____

3 ____

0+ ____

4+ ____

5 ____

1 ____

4 ____

3+ ____

____
NAME

____
DATE

____
LANGUAGE

Instructions, Phase 2

On the next page are pairs of speaking enter and exit proficiency ratings. Your task is to judge how difficult it would be for a typical LS language student to achieve each exit proficiency score given its paired enter proficiency score. You are to express this difficulty by assigning a number to each pair. The first pair of ratings, 1+ and 3, is to be called "10." Thereafter, you are to assign numbers proportional to your subjective impression of this first pair of ratings. For example, if you feel a particular pair of ratings is twice as difficult to achieve as the 1+ and 3 pair, assign to it a number "20." If you judge another pair to be one-fifth as difficult, call it "2," and so forth. Please do not restrict your response range. Use numbers as large or as small as you feel are necessary, including those less than "1" (fractions or decimals) if they are appropriate. Base all of your judgments on the same foreign language you chose in Phase 1. Please make note of this language at the bottom of the list.

SPEAKING

Enter	Exit	Difficulty
1+	3	10
3	4+	
0+	3	
2+	3+	
0	0+	
0	1	
1+	4	
2	3+	
1	2	
2	4	
0+	2	
2	2+	
1	5	
3	3+	
0	2	
2	4+	
0+	1+	
1+	2	
1	1+	
2	5	
0	3	
3+	4+	
1+	3+	
2+	4+	
2	3	
0	4	
0+	2+	
3+	5	
2+	4	
4	5	
3+	4	
0	4+	
3	4	
4	4+	
1+	4+	
0+	3+	
1	4	
1+	2+	
0	1+	
1+	5	

(Continued on page 252)

0+	4+	________
4+	5	________
0	3+	________
3	5	________
0+	5	________

1	3+	________
2+	5	________
1	2+	________
0	5	________
2+	3	________

1	4+	________
0+	4	________
0	2+	________
1	3	________
0+	1	________

Name: ______________________ Date: ______________ Language: _______________

References

Chomsky, N. *Language and Mind*. New York: Harcourt, Brace, Jovanovich, 1968.

Ekman, G. "A Simple Method for Fitting Psychophysical Power Functions." *Journal of Psychology*, 51 (1961): 343-50.

Mashhour, M. "On the Validity of Scales Derived by Ratio and Magnitude Estimation Methods." Psychological Laboratory, University of Stockholm. Technical Report No. 105, 1961.

Stevens, S. S. "A Scale for the Measurement of a Psychological Magnitude, Loudness." *Psychological Review*, 43 (1936): 405-16.

____________. "The Measurement of Loudness." *Journal of the Acoustical Society of America*, 27 (1955): 815-29.

____________. "The Psychophysics of Sensory Function." *American Scientist*, 48 (1960): 226-253.

____________. "To Honor Fechner and Repeal His Law." *Science*, 133 (1961): 80-86.

____________. "A Metric for the Social Consensus." *Science*, 151 (1966): 530-41.

Torgerson, W. S. *Theory and Methods of Scaling*. New York: Wiley, 1958.

SETTING STANDARDS OF SPEAKING PROFICIENCY

Samuel A. Livingston

Educational Testing Service

SETTING STANDARDS OF SPEAKING PROFICIENCY

Samuel A. Livingston

In our society we set standards for all kinds of things. The Food and Drug Administration sets standards for the purity of food products. The Environmental Protection Agency sets standards for the cleanliness of automobile exhaust fumes. And the New Jersey Department of Education sets standards for the speaking proficiency of teachers--in particular, teachers of English as a second language (ESL) and teachers of Spanish-English bilingual classes. A standard is simply an answer to the question: "How good is good enough?" Any answer to this question must involve judgment. Therefore, anyone who sets out to do a standard-setting study must answer four basic questions:

1. What type of judgments will enter into the standard-setting process?

2. Who will make those judgments?

3. How will the judgments be collected?

4. How will the judgments be used to determine the standard?

The purpose of this paper is to show how each of these four questions was answered in a standard-setting study conducted for the New Jersey Department of Education by Educational Testing Service. The Department of Education uses the Language Proficiency Interview (LPI) as a measure of speaking proficiency in certifying persons as eligible to teach ESL and Spanish-English bilingual classes. The standard-setting study was intended to help the Department decide what interview score level to establish as the minimum for certification for these teaching positions.

Of the four basic questions listed above, the first question--what type of judgments to use--is the most basic. In the case of the LPI, there are at least two ways to answer the question. One way is to use judgments made on the basis of the written statements that express the meanings of the various interview score levels. Another way is to use judgments of the actual interview performances of persons applying for certification. As the semanticists like to remind us, the word is not the thing; the written description of performance is not the performance itself. Therefore, we (that is, researchers from Educational Testing Service and administrators from the Department of Education) decided to base the standard-setting on judgments of the actual interview performances of individual candidates for certification: judgments of each speaker's proficiency as adequate or not adequate for the job in question (bilingual or ESL teacher).

The second question--whose judgments to use--depends partly on the types of judgments to be used. Our main concern was to choose a group of judges who would be representative of the population of persons qualified to judge a candidate's speaking proficiency as being adequate or inadequate for the job of a bilingual or ESL teacher. The Department of Education recruited three groups of judges, one group for each of three types of judgment:

1. English-language proficiency for ESL certification
2. English-language proficiency for bilingual certification
3. Spanish-language proficiency for bilingual certification

The judges were all experienced teachers (and in many cases also supervisors of teachers) of ESL or Spanish-English bilingual classes.

Collecting the Data

The third question--how to collect the data--involved a number of specific decisions. Considerations of scientific method entered into these decisions, as did administrative considerations. One important question of research design was how long a segment of each interview to present to the judges. Since the amount of time the judges could devote to the study was limited, we had to make a trade-off between two important considerations: getting a valid judgment of each interview presented and getting judgments of an adequate number of interviews at each score level. From a statistical point of view, if the total listening time is limited, the segments should be of the shortest length that will allow a meaningful judgment, so as to permit the judging of as many different interviews as possible. We decided to use five-minute segments, which enabled us to get judgments of twenty different interviews. (On the basis of our experience with this study, we now believe the judges could have made meaningful judgments of segments much shorter than five minutes.)

A related question is how to select the segment of each interview to present for judging. Experience with the LPI suggests that the portion of the interview that yields the most information about the examinee's strengths and weaknesses begins about thirty seconds after the opening of the interview. The opening thirty seconds usually consist of conventional greetings and simple introductory questions. During the following five minutes--the portion used in the study--the interviewer typically asks questions aimed at exploring the examinee's command of verb tenses and ability to communicate on several topics: personal and family background, personal activities and interests, teaching assignments, classroom activities, philosophies of education, and so on.

Another important question is the range of score levels to be represented in the study. Reducing the number of score levels allows more interviews at each of the remaining levels, but it is important not to exclude any levels that might turn out to be near the standard. We eliminated levels 0 and 0+ and level 5, assuming that almost no level 0 or 0+ interviews would be judged adeqjate and that most level 5 interviews would be judged adequate. This decision enabled us to present three interviews at all but one of the remaining seven score levels. Level 4+ was represented by only two interviews, instead of three.

We decided to use the same sample of English-language interview segments for both the ESL and English-bilingual judging. This decision enabled us to make direct comparisons between the ESL and English-bilingual judgments. It also simplified the data collection procedure.

To avoid "sequence effects"--systematic trends in the sequence of score levels of the interview segments that might bias the judgments--we used the following procedure. First, we divided the twenty interview segments into three subsamples so that each subsample contained an interview segment at every score level (with one exception: level 4+ was not represented in the last subsample). We then randomized the order of the score levels in each subsample, using a different random sequence for each subsample. This procedure produced the following sequence of score levels: 2, 4, 4+, 3, 1+, 2+, 3+, 3, 4+, 3+, 2+, 2, 4, 1+, 2+, 4, 2, 3+, 1+, 3. We used the same sequence for both the English-language interviews and the Spanish-language interviews.

The actual judging took place at the language laboratory of Rider College in Trenton, New Jersey. Eight ESL judges, eleven English-bilingual judges, and eleven Spanish-bilingual judges participated. The judges received instructions emphasizing that their task was to judge whether the speaking proficiency of the person being interviewed in each segment was "at least minimally sufficient for this person to function adequately" in the relevant teaching job. The judges listened to the taped interview segments through earphones at individual listening booths. They were instructed not to communicate with each other during the judging process or to give any audible or visible reaction to the interview segments.

Analysis of the Data

Our data analysis was intended to take the information contained in the individual judgments and summarize it in such a way that it would be as useful as possible for setting standards. Therefore, we tried to present the results of the judging in a way that would answer the question: "Given a candidate's interview score, what is the probability that the candidate's actual speaking proficiency would be judged acceptable?" Another way to express this question is to ask, "If all interviews at a given score level were judged by all possible judges, what percentage of the resulting judgments would rate the candidate as acceptable?" We sought to answer this question for the English speaking proficiency of ESL teachers, the English speaking proficiency of Spanish-English bilingual education teachers, and the Spanish speaking proficiency of Spanish-English bilingual education teachers.

English as a Second Language. The results of the judging of the English-language interview segments by the eight ESL judges are shown in Table 1 and presented graphically in Figure 1. Table 1 shows what percentage of the judges rated each interview segment acceptable, as well as the average of these percentages for all the interview segments at each LPI score level. For example, of the three interview segments at score level 3, the first was considered acceptable by 25 percent of the judges; the second, by 88 percent; and the third, by 100 percent. The average of these three percentages is 71 percent. This average can be interpreted as an estimate of the probability that a randomly selected level 3 interview would be rated as acceptable by a judge selected at random from the population of all possible ESL judges. Note that in Table 1 these estimates increase steadily from zero at level 1+ to 100 percent at level 4+.

The fact that fourteen of the twenty interview segments were judged acceptable either by none of the ESL judges or by all of the ESL judges indicates a high degree of consistency. In fact, for seventeen of the twenty segments, at least seven of the eight ESL judges were in agreement, even though they made their judgments independently, without any communication with each other.

Figure 1 provides a graphic presentation of the information in Table 1. The dots represent the percentages of acceptance for the individual interview segments. Horizontal lines have been drawn at 0, 50, and 100 percent to make the graph easier to read. For the same reason, vertical lines have been drawn to connect the dots representing interview segments at each score level. The average percentage of acceptance at each score level is indicated by a short horizontal line. Notice that the average percentage of acceptance rises steadily from level 1+ to level 4+ in such a way as to suggest a smooth curve. If such a curve were drawn on the graph, it would cross the dashed line indicating 50 percent acceptance somewhere between level 2+ and level 3.

English-Bilingual. Table 2 and Figure 2 present the results of the judging of the English language tapes by the English-bilingual judges. These judges also appear to have been quite consistent in their evaluations (though not quite as consistent as the ESL judges). The average percentage of acceptance of the English language interviews is consistently higher for the English-bilingual judges than for the ESL judges. This result suggests that the teaching of English as a second language requires a higher level of English-language speaking proficiency than does the teaching of bilingual education classes.

The average percentage of acceptance by the English-bilingual judges (like that by the ESL judges) increases steadily with increasing score levels, from 21 percent at level 1+ to 100 percent at levels 4 and 4+. A smooth curve connecting these points in Figure 2 would cross the line representing 50 percent acceptance slightly above score level 2 (rather than between 2+ and 3, as was the case for the ESL judgments).

TABLE 1

English as a Second Language
(8 judges)

Interview Score Level	Percentage of Judges Accepting Interview Segment			
	Tape 1	Tape 2	Tape 3	Average
1+	0	0	0	0
2	0	0	12	4
2+	0	100	12	38
3	25	88	100	71
3+	62	100	100	88
4	100	88	100	96
4+	100	100	--	100

FIGURE 1

Acceptability Judgments for English as a Second Language

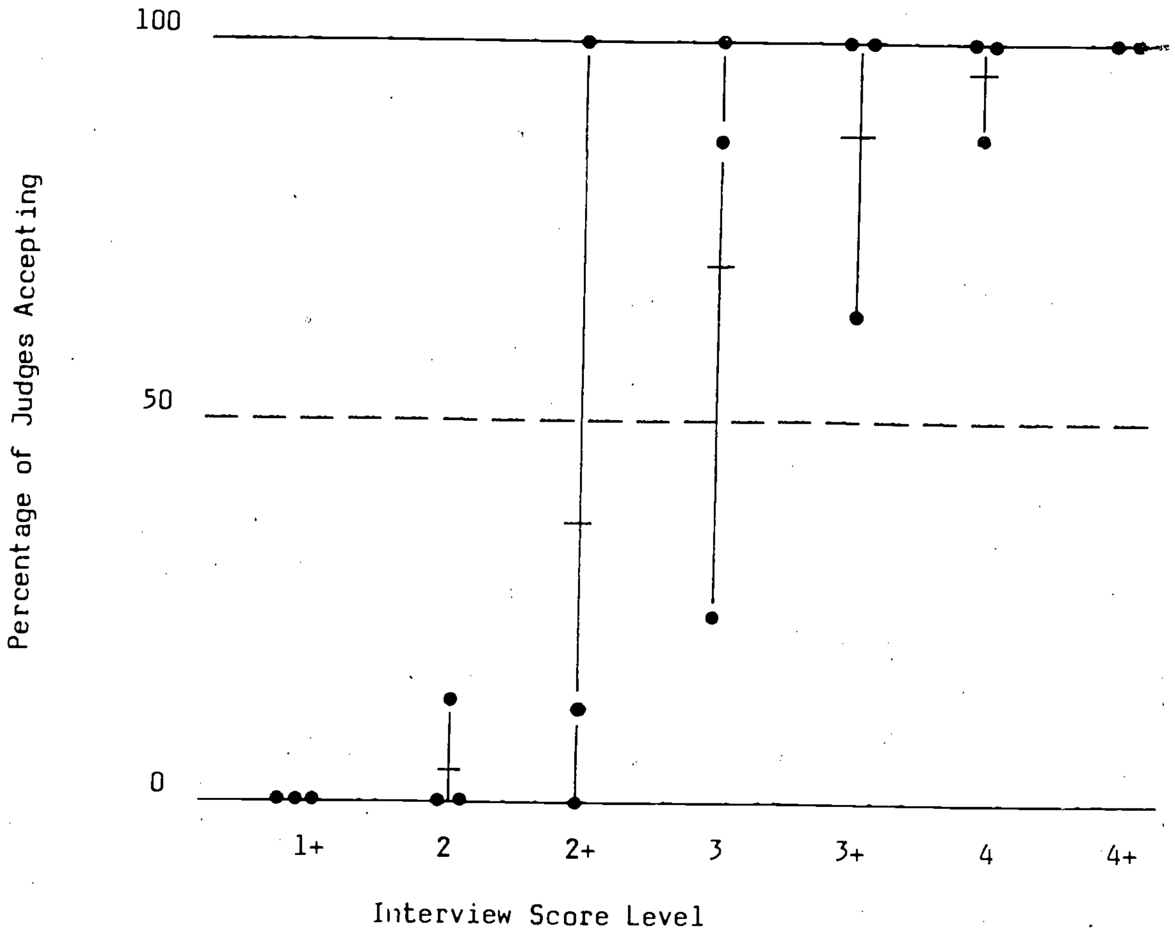


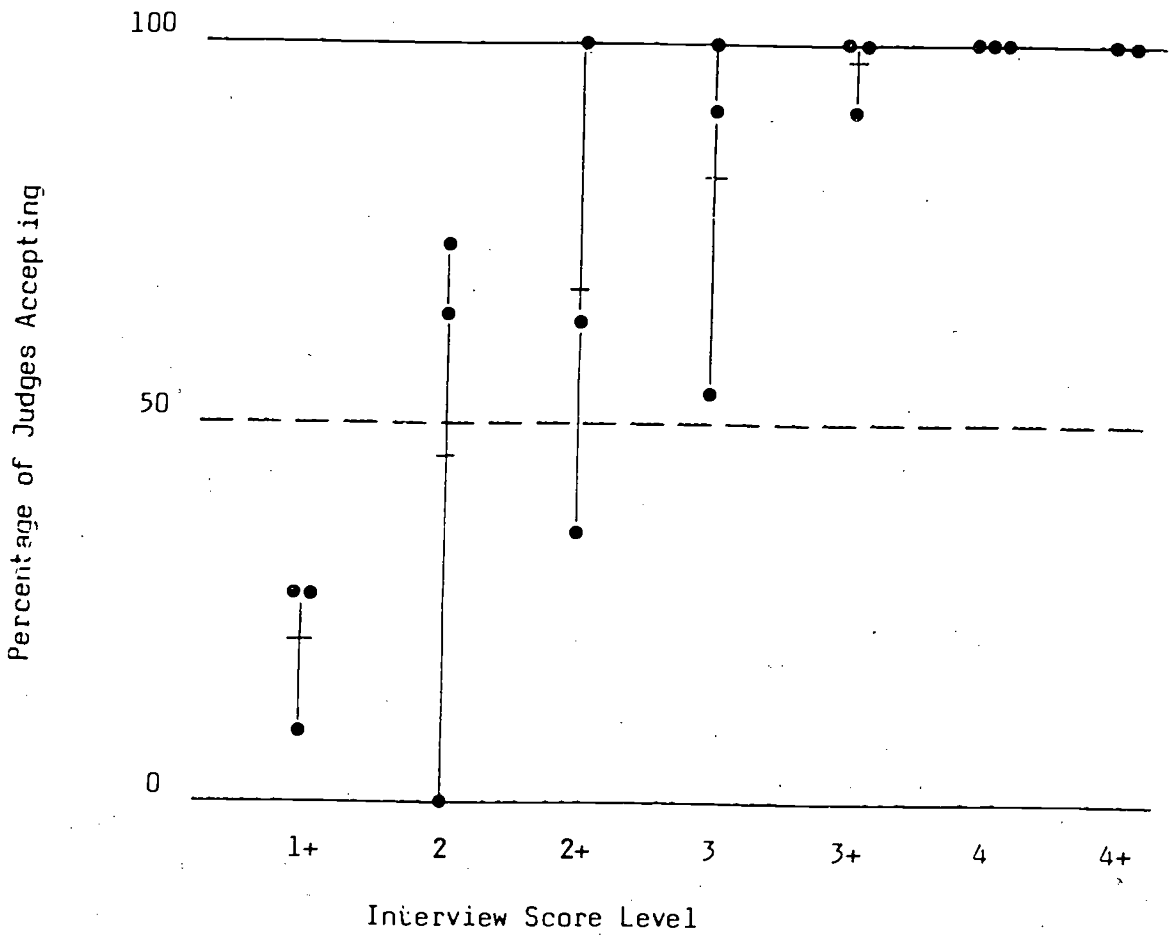
TABLE 2

English Component of Bilingual Education
(11 judges)

Interview Score Level	Percentage of Judges Accepting Interview Segment			
	Tape 1	Tape 2	Tape 3	Average
1+	27	27	9	21
2	0	73	64	45
2+	36	100	64	67
3	55	100	91	82
3+	100	100	91	97
4	100	100	100	100
4+	100	100	--	100

FIGURE 2

Acceptability Judgments for
English Component of Bilingual Education



Spanish-Bilingual. Table 3 and Figure 3 show the results of the judging of the Spanish-language tapes by the Spanish-bilingual judges. These judges appear to have been slightly less consistent in their evaluations than the English-bilingual judges. However, at least ten of the eleven judges agreed on eleven of the twenty interview segments, and a clear majority of the judges were in agreement on all but one of the interview segments.

The results of the judging of the Spanish-language interview segments differ in one obvious way from the results of the judging of the English-language segments: the average percentage of acceptance does not rise steadily from one score level to the next, but shows a somewhat inconsistent pattern between levels 2+ and 4. These inconsistencies are probably the result of sampling variability in the small number of interview segments presented for judging. A desirable approach in such a situation would be to get judgments of several additional interview segments at these levels. This approach, however, would require reconvening the Spanish-bilingual judges for a further judging session.

One way to deal with these fluctuations in the observed data is by means of a statistical technique known as "smoothing." The rationale for the use of smoothing with these data is the assumption that if we could somehow get judgments of all possible interviews at each score level, the average percentage of acceptance would increase steadily across the score levels. Thus, if a graph similar to Figure 3 were drawn on the basis of judgments of all possible interviews, the points representing the average percentage of acceptance would follow a smooth rising curve, as they do in Figures 1 and 2. The purpose of smoothing is to provide a statistical estimate of that curve on the basis of the available data. This estimated curve is shown in Figure 3. Smoothing improves the estimation at each score level by making use of information contained in the data from the adjacent score levels. The smoothing formula we used can be stated in words as follows: for each score level, the estimated (smoothed) percentage of acceptance is given by:

one-half of the pe-centage of acceptance at that score level, plus

one-fourth of the percentage of acceptance at the next lower score level, plus

one-fourth of the percentage of acceptance at the next higher score level.

The smoothed averages are an improvement over the actual observed averages, in the sense that they can be expected to provide a better estimate of what the averages would have been had the judging session included a very large number of interview segments at each score level.

TABLE 3

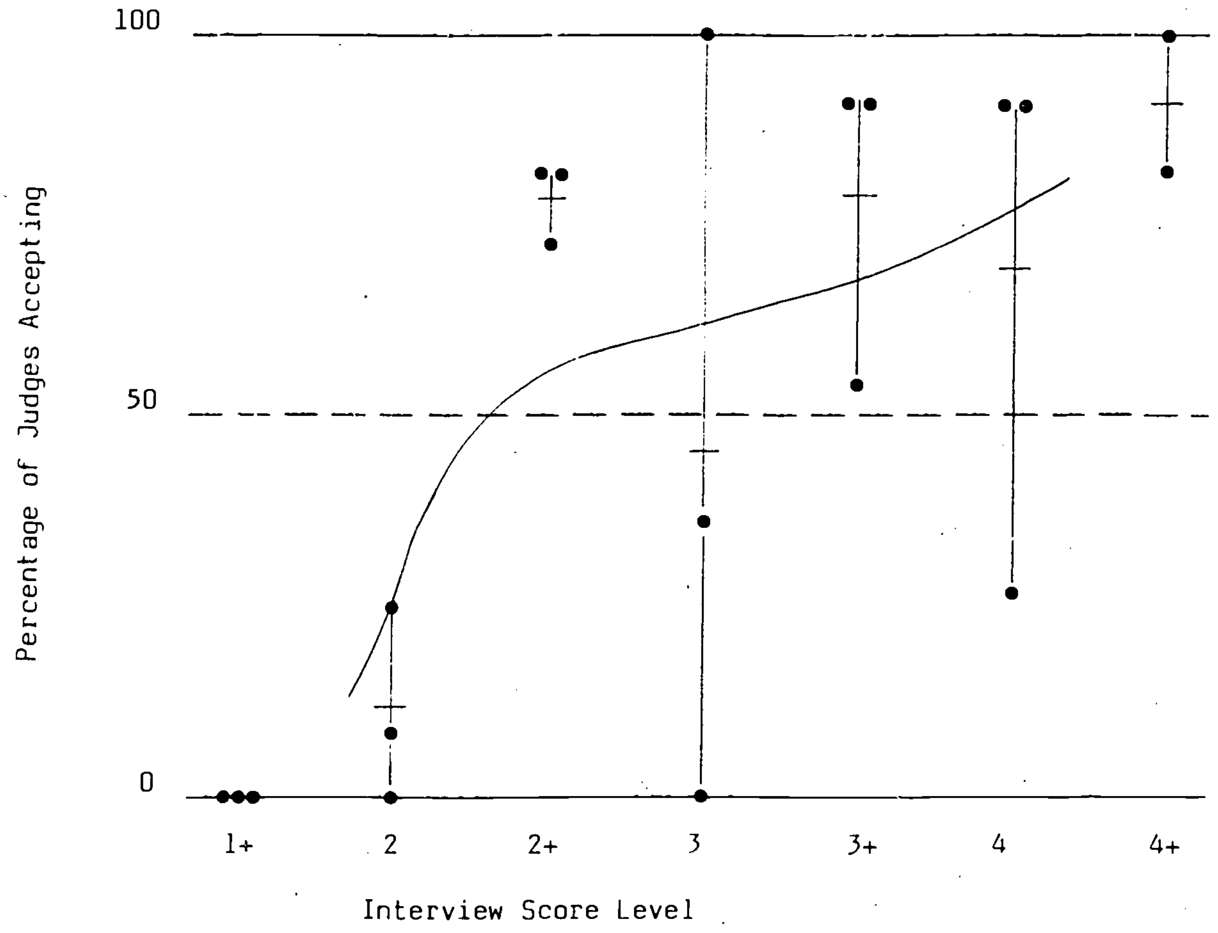
Spanish Component of Bilingual Education
(11 judges)

Interview Score Level	Percentage of Judges Accepting Interview Segment				
	Tape 1	Tape 2	Tape 3	Average	
				Actual	Smoothec
1+	0	0	0	0	*
2	9	27	0	12	26
2+	73	82	82	79	54
3	100	36	0	45	62
3+	91	91	55	79	68
4	91	27	91	69	77
4+	100	82	--	91	*

*The smoothing formula used does not provide for computation of smoothed values at the highest and lowest levels.

FIGURE 3

Acceptability Judgments for
Spanish Component of Bilingual Education



The smoothed average percentages of acceptance by the Spanish-bilingual judges are shown in the last column of Table 3. These percentages are lower than the corresponding percentages for the English-bilingual interviews at every score level. However, the curve in Figure 3 crosses the line representing 50 percent acceptance at a point between interview score levels 2 and 2+, as is the case for the English-bilingual judging. These results suggest that the teaching of Spanish-English bilingual classes requires a degree of Spanish-language proficiency that is at least as high as the degree of English-language proficiency required, and possibly somewhat higher.

Setting the Standard

The research study provides an estimate of the relationship between a speaker's interview score and the probability that the speaker's proficiency will be judged adequate. It does not tell the decision maker how to use this information to set a standard. One way to proceed is to set the pass/fail cutoff for interview scores at the point where the probability of acceptance equals 50 percent. This choice has a simple rationale: speakers with interview scores below the cutoff are more likely to be judged unacceptable than they are to be judged acceptable, while the reverse is true for speakers with interview scores above the cutoff.

Any decision based on less than perfect information involves the possibility of error. In the case of a pass/fail decision about a speaker whose interview score is known, there are two types of errors: passing a speaker who would have been judged inadequate, and failing a speaker who would have been judged adequate. The rationale for setting the pass/fail cutoff at the score that corresponds to 50 percent acceptance is based on the implicit assumption that these two types of errors are equally serious. But what if they are not equally serious? For example, what if it is twice as serious an error to pass an inadequate speaker as to fail an adequate speaker? Obviously, in this case, the cutoff should be somewhat higher than the score that corresponds to a 50 percent probability of acceptance, but how much higher?.

Statistical decision theory (which, at its simplest levels, is really common sense expressed in mathematical language) provides the following answer: If it is twice as serious an error to pass an inadequate speaker as to fail an adequate speaker, we can tolerate two errors of the second kind (failing a person who should pass) for every error of the first kind (passing a person who should fail). Therefore, we should raise the cutoff to the interview score level at which there are twice as many adequate speakers as inadequate speakers. This is the score level that corresponds to a two-thirds (or 67 percent) probability of acceptance. At any interview score above this cutoff, the adequate speakers will outnumber the inadequate speakers by more than two to one, so we will do more harm by failing the adequate speakers than by passing the inadequate speakers at that score level. At any interview score below the cutoff, the number of adequate speakers is less than twice the number of inadequate speakers, so we will do more harm by passing the inadequate speakers than by failing the adequate speakers at this level.

The standard-setting process, therefore, involves two kinds of judgment. The first is the judgment of speakers' proficiency as adequate or inadequate. The second is the judgment of the relative seriousness of the two types of possible errors. These two kinds of judgment do not have to be made by the same persons, and often they will not be, since different kinds of competence are involved. The first kind of judgment requires the ability to recognize adequate and inadequate performance; the second requires the ability to evaluate the consequences of adequate and inadequate performance.

Summary

The New Jersey LPI study is an example of a more general procedure for conducting an empirical standard-setting study. This general procedure can be described as follows:

1. Determine the measure of performance for which the standard is to be set. In general terms we can call this measure the test score. In the New Jersey study it was the Language Proficiency Interview score.

2. Determine the type of performance that will serve as the basis for judging a person's proficiency as adequate or inadequate. In general terms we would call this performance the criterion performance. The criterion performance in the New Jersey LPI study was a portion of the interview itself.

3. Identify a population of persons qualified to judge examples of the criterion performance as adequate or inadequate. Select a sample of these persons to serve as judges.

4. Identify the population of persons taking the test for which a standard is to be set and obtain their test scores. Select a sample of these examinees, making sure the range of their test scores is broad enough to include both the lowest and the highest scores that might conceivably be selected as the standard.

5. Obtain judgments of the examinees' criterion performances by the judges.

6. Analyze the data provided by these judgments to estimate the probability that an examinee's criterion performance will be judged adequate, as a function of the examinee's test score.

These six steps make up the empirical study. Two remaining steps complete the standard-setting procedure.

7. Determine the relative seriousness of the two types of possible errors: passing an examinee whose criterion performance is inadequate and failing an examinee whose criterion performance is adequate.

8. Set the standard at the test score level that results in an equal risk of the two t/pes of possible errors, weighted by heir seriousness in the particular decision-making situation for which a standard is to be set.