1986

Evaluating oral proficiency and communicative competence in foreign languages

Paul Allen Beaufait

The University of Montana

Let us know how access to this document benefits you.
Follow this and additional works at: https://scholarworks.umt.edu/etd

Recommended Citation
Beaufait, Paul Allen, "Evaluating oral proficiency and communicative competence in foreign languages" (1986). Graduate Student Theses, Dissertations, & Professional Papers. 7838.
https://scholarworks.umt.edu/etd/7838

This Thesis is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.
COPYRIGHT ACT OF 1976

THIS IS AN UNPUBLISHED MANUSCRIPT IN WHICH COPYRIGHT SUB-
SISTS. ANY FURTHER REPRINTING OF ITS CONTENTS MUST BE APPROVED
BY THE AUTHOR.

MANSFIELD LIBRARY
UNIVERSITY OF MONTANA
DATE: 1986
EVALUATING ORAL PROFICIENCY AND COMMUNICATIVE COMPETENCE IN FOREIGN LANGUAGES

By

Paul Allen Beaufait

Presented in partial fulfillment of the requirements for the degree of

Master of Arts

University of Montana

1986

Approved by

Chairman
Board of Examiners

Dean,
Graduate School

Date
This study first examines three instruments available in the United States to measure oral proficiency in foreign languages: the Oral Interview (OI), the Test of Spoken English (TSE) and the Advanced Placement French Language (APFL) Speaking Subsection. Then, it presents findings related to language competence and, specifically, to testing strategies. This study concludes by summarizing oral proficiency applications and implications.

Each instrument is described critically in terms of form, reliability and validity. First, two variants of the OI are compared: the United States government's latest version and the American Council on the Teaching of Foreign Languages/Educational Testing Service (ETS) academic version. The OI, although sophisticated and time consuming, is a face-valid measure of individuals' oral proficiency. The other two instruments are standardized tests administered to groups. Both the College Entrance Examination Board's APFL examination and the TSE were developed by ETS, which explains the similarities between the two. Since current validity data are lacking for these two instruments, they must be viewed as limited in valid applications. It is as group test alternatives to the OI that these two merit attention.

Of two hypotheses, indivisible language competence and partially divisible competence, the first is generally supported by research in language testing. Groups developing oral tests, especially abroad, have extended both contexts and scoring criteria to encompass more communicative components, and have begun identifying individuals' communicative needs. This is a necessary step toward communicative competence evaluation in foreign languages. The problem of valid group evaluation remains.

Applications for oral proficiency evaluation, in academia, include pre-, in-, and post-course measurements, and new teacher certification requirements. Curricular and methodological development, and tester training, must accompany evaluation, if oral proficiency is to be successfully taught and tested. Non-academic needs include the need for qualified language speakers, the need for measurements for and in jobs.
**CONTENTS**

<table>
<thead>
<tr>
<th>CHAPTER.</th>
<th>Section.</th>
<th>INTRODUCTION</th>
<th>1.00 ORAL PROFICIENCY MEASUREMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>Problem</td>
<td></td>
<td>6.00</td>
</tr>
<tr>
<td>0.11</td>
<td>Current Needs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>Study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.21</td>
<td>Resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>Progress</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>INTRODUCTION NOTES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>ORAL PROFICIENCY MEASUREMENT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.01</td>
<td>The Oral Interview</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.02</td>
<td>The Test of Spoken English (TSE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.03</td>
<td>The Advanced Placement Examination in French Language (APFL): Speaking Subsection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.04</td>
<td>Current Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.10</td>
<td>The Oral Interview</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>1.11.00</td>
<td>General Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.11.01</td>
<td>Oral Interview Format</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.11.10</td>
<td>Oral Interview Rating Scales</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.11.11.0</td>
<td>Ratings on the Government Scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.11.11.1</td>
<td>Levels: Generic Descriptions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.11.12</td>
<td>Academic Scale Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.11.13</td>
<td>Ratings on the Academic Scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.11.14</td>
<td>Academic/Government Scale Comparisons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.11.15</td>
<td>Language Specific Descriptions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.11.2</td>
<td>Function in the Oral Interview Context</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.12.0</td>
<td>Oral Interview Reliability</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>1.12.1</td>
<td>Rater Reliability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.12.2</td>
<td>ACTFL/ETS Workshops</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.13.0</td>
<td>Oral Interview Validity</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>1.13.1</td>
<td>Criterion Related Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.13.20</td>
<td>Construct Validity</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
1.31.20 Speaking Subsection Function
1.31.21 Directed Responses
1.31.22.0 Storytelling
1.31.22.1 Illustration
1.31.22.2 Scoring Criteria
1.32.0 APFL Reliability
1.32.1 Free Response Scoring
1.32.2 Final Grade
1.33.0 APFL Validity
1.34.0 APFL Critique
1.34.1 Consanguinity: Speaking Cousins
1.34.2 Scoring
CHAPTER 1 NOTES

2.00 ORAL PROFICIENCY AND COMMUNICATIVE COMPETENCE EVALUATION

2.10 Oral Proficiency Measurement
2.11.0 An Intensive English Language Program Study
2.11.1 Discrete Point vs. Cloze Testing
2.12.0 French, Spanish, and Russian Proficiency Measures
2.12.1 French and Spanish Guide
2.12.2 Russian Proficiency Handbook
2.13 Indivisible Competence
2.14.0 Native Speaker Studies
2.14.1 An Accented English Proficiency Study
2.14.2 Native Speaker Error
2.15.0 Success Factors
2.15.1 Behavioral and Attitudinal Correlates
2.15.2 Self Reported Data
2.15.3 Integrative vs. Instrumental Motives
2.20 Toward Communicative Competence
2.21.0 Negotiating Understanding
2.21.1 Communicative Features
2.21.2 Implications
2.22.0 Sociolinguistic, Interpersonal and Pragmatic Variables
2.22.1 Setting
2.22.2 Interlocutors
2.22.3 Communicative Outcome
2.23 Communicative Competence Caveat
2.24 Dutch Developments
<table>
<thead>
<tr>
<th>CHAPTER.</th>
<th>Section.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.11.14</td>
<td>Figure 1</td>
</tr>
</tbody>
</table>
0.00 INTRODUCTION

0.10 Problem

It is evident, to most foreign language instructors, that the progress learners make in acquiring foreign languages requires methodic evaluation. This holds true for the proficiency that learners may demonstrate in listening, speaking, reading and writing foreign languages. Though numerous tests now exist to evaluate proficiency in the first three of these: listening, reading and writing, oral proficiency requires special attention.

Paper-and-pencil tests, standard instruments such as the Test of English as a Foreign Language (TOEFL) or the College Level Examination Program French Test which are used for admission and placement purposes,\(^1\) do not directly measure speaking proficiency. As a result, the major burden of evaluating oral proficiency falls on the foreign language instructor who may find that an informal office conversation with a language learner in the target language yields greater insights to the learner's speaking ability than do standard test scores.\(^2\)

Because language learners need the advice of instructors with regard to appropriate courses of study,
learners, too, must share the burden imposed by the lack of oral proficiency measurements. It behooves these learners to verify their speaking proficiency before travelling to a country where the target language is spoken. Similarly, graduate teaching assistants who have to teach in a foreign language must communicate with their students, as all adult language learners must with their communicative partners.

"Too often speaking has gone untested (Lowe 1984, 39)." So language instructors need a means to effectively measure speaking proficiency and communicative competence, to diagnose weakness and to predict success in these two areas.

Current Needs

Pardee Lowe, Jr. and Judith E. Liskin-Gasparro (1983) argue that there is substantial need for oral proficiency measurement. What foreign language teachers need are techniques that will enable them to measure second language speaking proficiency as accurately as they can measure other skills. Oral proficiency measurement has applications:

... in those situations where proficiency testing is in order, e.g., placement testing; [testing before] and after intensive language training; testing prior to and after living abroad; testing at the end of a major sequence of high school of college courses; testing for course credits awarded for proven proficiency ... ; testing for suitable language ability for teacher certification for high school teachers and graduate teaching assistants (Lowe and Liskin-Gasparro 1983, 3).
Study

In this study I investigate different methods of evaluating oral proficiency and communicative competence in foreign languages. In order that other language instructors may benefit from this investigation, I include references and suggestions applicable to various languages and proficiency levels. Both language instructors and the adult language learners should find them valuable in meeting their specific measurement needs and language acquisition goals.

Resources

Considering the rate of growth in not only the number of measurement instruments but also of evaluation strategies, I utilize the most up to date resources available. I have selected a representative sample from the instruments available in the United States and have limited the sources cited, in most instances, to the last five years (1981-1985). Wherever possible, I have selected sources dating from 1983 or later.

Progress

Direct measurement of oral proficiency ranges among the most rapidly developing spheres of interest to foreign language instructors present and future, as well as to adult language learners. So the current state of oral
proficiency testing calls for professional and scholarly attention. In the future, I hope to see foreign language instructors make the most of the oral proficiency measurement techniques that have been and are being developed.
INTRODUCTION

NOTES

1University of Montana, 1984–85 Catalog, 592 (September 1984): 24–5; and conversations with Drs. O. W. Rolfe and Anthony Beltramo, French and Spanish Sections, Department of Foreign Languages and Literatures, University of Montana, Missoula, MT: Winter-Spring 1985.

1.00 ORAL PROFICIENCY MEASUREMENT INSTRUMENTS

In this chapter, I describe three measurement instruments and evaluate them in terms of reliability and validity. A critique which addresses questions and concerns of foreign language instructors interested in classroom applications then follows the description and evaluation of each instrument.

The instruments, in order of presentation, are:

1. the Oral Interview
2. the Test of Spoken English
3. the Advanced Placement Examination in French Language Speaking Subsection

1.01 The Oral Interview

Because of the prominence of the Oral Interview in the oral proficiency testing literature, and with recent efforts by academic and professional organizations in conjunction with the United States government to adapt the interview for academic measurement purposes, it is appropriate to examine the Oral Interview first in this chapter.

The Oral Interview was developed by the United States government for the purpose of evaluating the speaking proficiency of military and foreign service personnel. Since then, both industry and the academic community have
shown an interest in applications of the Oral Interview. The American Council on the Teaching of Foreign Languages (ACTFL) and the Educational Testing Service (ETS) cooperated in adapting the Defense Language Institute, the government's, Oral Interview rating scale for academic uses (Heidinger 1983).

Lowe and Liskin-Gasparro (1983, 2) cite the following reasons for Oral Interview adaptation:

Second language teachers at the high school and post secondary levels realized that the [government] scale would need some adjustment to adequately measure their student's proficiency [because] the lower end of the scale (0, 0+, 1, 1+) was not sensitive enough to register observable differences in the proficiency among students, and the upper end of the scale would seldom be appropriate for the skill levels that students could be expected to achieve.

So ACTFL, ETS and the Interagency Language Roundtable, a group of government agencies including the Departments of Education and State, developed a revised rating scale. ACTFL and ETS now teach workshops on the result of their efforts, the ACTFL/ETS or academic rating scale, to those persons interested in testing the oral proficiency of high school and post secondary students (Lowe and Liskin-Gasparro 1983, 2). Consequently the Oral Interview has applications in academia as well as in government.

A final reason for giving priority to the Oral Interview is the fact that ETS used the Oral Interview as a standard of comparison in developing the Test of Spoken
English (Clark and Swinton 1980, iii). So, in this chapter, I treat the Oral Interview before the Test of Spoken English.

1.02 The Test of Spoken English (TSE)

I have included the TSE here for consideration by those who are interested in evaluating the speaking proficiency of groups rather than individuals, because the TSE format may prove more readily adaptable to classroom applications than Oral Interview techniques.

To colleges and universities concerned with the qualifications of foreign Graduate Teaching Assistants (GTAs), ETS offers a group-test alternative to the Oral Interview, the TSE, which institutions may use to measure GTAs' English speaking proficiency (TSE Manual 1982, 5). The TSE also may replace, or simply supplement, the Test of English as a Foreign Language (TOEFL) as the standard instrument for evaluating college-bound foreign students' English proficiency, especially of those who like GTAs must speak English proficiently.

1.03 The Advanced Placement Examination in French Language (APFL) Speaking Subsection

In the third section of this chapter I describe and evaluate another group-test alternative to the Oral Interview, the College Board's APFL Speaking Subsection.
This treatment of the APFL Speaking Subsection complements that of the TSE in two respects.

First, both the APFL Speaking Subsection, in a Romance language, and the TSE, in a Germanic language, were developed by ETS. While the former is only part of a larger instrument and the latter an instrument in and of itself, there are a number of striking similarities between the two.

Second, the Advanced Placement Program has compared the performance of students on the APFL with that of college students who have completed six semesters of French at the college level. Although the APFL examination is intended for qualified students near the end of their secondary school training (CEEB 1984, 2), and the TSE for college graduates, the two populations' proficiency is comparable. APFL candidates' proficiency roughly approximates that of foreign students who have studied English as a Second or Foreign language, and who intend to take college courses taught in English.

Even though it is difficult to determine exact equivalence among programs in English as a Second or Foreign Language abroad and foreign language programs in United States high schools, foreign GTAs capable of teaching college level courses in English definitely exhibit advanced skill. Similarly, the intended populations for the TSE and the APFL differ in terms of provenence (foreign
vs. U.S.). Nonetheless they can be compared in terms of success in college level courses taught in a foreign language, what the APFL is intended to predict.

1.04 Current Reference

The academic and government Oral Interview rating scales available for the present investigation date from 1982 and 1983. Selected sources referring to the TSE date from 1980 to 1984. Finally, references to the APFL and to the Speaking Subsection include the May 1985 College Board description (CEEB 1984).

1.10 The Oral Interview

The Oral Interview (OI) serves as a technique for establishing oral proficiency (Heidinger 1983). Before moving on to examine recent stages in the ongoing development of the OI, I would like to describe this technique briefly.

1.11.00 General Description

The OI is a proficiency test that measures examinees’ speaking ability, the ability to use a language, the ability to create with the language which the examinee has acquired (Lowe 1984, 35-6). The OI may be used to evaluate Non-Indo-European language (Chinese for example) speaking proficiency in addition to Indo-European language speaking proficiency, Romance (French and Spanish) as well
as Germanic (English and German). Lowe and Liskin-Gasparro (1983, 1) suggest that this instrument is capable of measuring a wide range of speaking ability.

The purpose of the OI is to evaluate examinees' speaking ability relative to that of educated native speakers rather than to coursework and course material. Because the interview is not based on a particular curriculum, it differs from an achievement test (Lowe and Liskin-Gasparro 1983, 2). It also breaks with discrete-point testing tradition in that the OI approach requires examinees to demonstrate integrative and functional language ability (Lowe 1984, 38).

During the OI, one or more trained testers direct conversation with an examinee. The interview lasts from ten to forty minutes depending upon the proficiency that the examinee demonstrates in speaking on a variety of topics (Lowe and Liskin-Gasparro 1983, 1).

1.11.01 Oral Interview Format

The format of the interview calls for a warm-up, lasting two to three minutes, which is intended to put the examinee at ease and to allow the tester to determine roughly at which level the examinee is able to function. Then the tester thoroughly checks the examinee's proficiency in different contexts at the level predetermined in the warm-up.
Next the tester probes a higher level in various contexts until a linguistic breakdown occurs. In this phase of the interview the tester must devise conversation-evoking situations in which the examinee does more than simply respond to a series of questions. After the linguistic breakdown the tester guides the conversation back to a level where the examinee feels comfortable (Heidinger 1983). Speech samples elicited by testers in the OI are often recorded for later verification.

1.11.10 Oral Interview Rating Scales

Two rating scales are currently available in the United States. The government uses its own Interagency Language Roundtable (ILR) Rating Scale (formerly known as the Defense Language Institute, DLI, rating scale).³ Recently however, with government cooperation, ACTFL and ETS have developed a parallel academic scale.

In the following sections, I first describe ratings based on the government scale, then examine reasons for subsequent development of the ACTFL/ETS rating scale and, finally, compare the academic scale to the government scale.

1.11.11.0 Ratings on the Government Scale

After the interview, testers rate speech samples according to a predetermined scale. Their ratings range
from zero, for no practical ability to function in the language, to five, for ability equivalent to that of a well-educated native speaker.

In addition to those six level ratings, testers may assign a plus rating (0+, 1+, 2+, up to 4+). In order to receive a plus rating, examinees must demonstrate ability exceeding the requirements for a particular level yet fail to exhibit consistently higher level performance under probes. Consequently testers must be capable of distinguishing among eleven ranges of speaking proficiency.

OI testers thus evaluate examinees' overall speaking proficiency with respect to functional language ability exhibited in the interview. Each of the eleven proficiency ranges corresponds to a definition based upon real linguistic needs and behaviors (Lowe and Liskin-Gasparro 1983, 1).4

1.11.11.1 Levels: Generic Descriptions

Heidinger (1983) suggests that a language learner functioning at Level 1 is possibly capable of surviving for a day in the society or culture which uses the language. A speaker at Level 2 could function for an indefinite, short period of time in a foreign language situation. Heidinger (1983) also considers Level 2 a significant level of achievement for advanced placement students from high
schools, foreign language majors at universities and high school foreign language instructors.

According to Heidinger, speakers at Level 3 are able to think and handle abstractions in the language, while proven proficiency at Level 4 distinguishes those who have lived in the country where the language is spoken and who have used the language for a long time. Since speakers at Level 4 are almost bilingual, the United States government does not attempt to train speakers beyond Level 3.

Level 5 represents speaking proficiency equivalent to that of an educated native speaker who has received an education, probably beyond the secondary level, in a society or culture using the language on a daily basis. Heidinger goes so far as to say that even among native speakers it is quite possible to find proficiency no greater than Level 2+ or 3.

1.11.12 Academic Scale Development

ACTFL and ETS had been cooperating with the government in adapting the government proficiency level definitions to academic needs and expectations for several years when Heidinger reported, in 1983, that ACTFL was still using the government scale. At that time, the government had been using the OI for more years than I choose to report here.5

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Pardee Lowe, Jr. (1984, 32-41) also mentions changes in the lower end of the rating scale, which were desirable prior to 1981 for two reasons, a stricter distinction, first, between proficiency and achievement testing and, second, between integrative and discrete-point testing. These changes were accomplished in 1982, when ACTFL and ETS formulated Provisional Generic Language Descriptions for speaking proficiency. Although the 0I that he documents constitutes a mixture of both government and academic forms, Lowe focuses primarily on the ACTFL/ETS modifications.

At that time, the proficiency level descriptions were expanded, particularly in the zero to one range on the government scale, in line with more realistic expectations and attainable goals for high school and college students (Lowe and Liskin-Gasparro 1983, 2). This was because ACTFL and ETS sought greater sensitivity to proficiency on the lower portion of the government scale (Lowe 1984, 34).

1.11.13 Ratings on the Academic Scale

The 1982 ACTFL Provisional Generic Descriptions define three proficiency levels: Novice, Intermediate and Advanced, and include the omnibus category Superior. The Novice and the Intermediate levels embody three distinct subdivisions each: Low, Mid, and High. The designation Advanced Plus supplements the Advanced level. It seems
that college foreign language majors generally reach the Advanced or Advanced Plus level (Lowe 1984, 43-49).

Lowe (1984, 38) admits that achievement does play a part in the ACTFL Novice ranges, where it is more a matter of language learners producing memorized material in a suitable context than of creating with the language. Yet, at the Intermediate Level and beyond, it is clearly a matter of proficiency, as opposed to achievement, since most of the questions posed are not based on classroom material.

1.11.14 Academic/Government Scale Comparisons

In order to illustrate the differences between the Government and Academic scales, I have adapted a side-by-side table comparing the two (see Figure 1). The ACTFL/ETS scale shown introduces further distinctions at government Levels 0 and 1. They are Novice and Intermediate Low, respectively. ACTFL/ETS Advanced equates with Level 2, Advanced Plus with Level 2+. At government Level 3 and above, the ACTFL/ETS scale makes no further distinctions. This means that speakers with government proficiency ratings from three to five all would fall into the Superior category if rated on the academic scale.
<table>
<thead>
<tr>
<th>ILR</th>
<th>ACTFL/ETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4+</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3+</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>SUPERIOR</td>
</tr>
<tr>
<td>2+</td>
<td>ADVANCED PLUS</td>
</tr>
<tr>
<td>2</td>
<td>ADVANCED</td>
</tr>
<tr>
<td>1+</td>
<td>INTERMEDIATE HIGH</td>
</tr>
<tr>
<td>1</td>
<td>INTERMEDIATE MID</td>
</tr>
<tr>
<td>1</td>
<td>INTERMEDIATE LOW</td>
</tr>
<tr>
<td>0+</td>
<td>NOVICE HIGH</td>
</tr>
<tr>
<td>0</td>
<td>NOVICE MID</td>
</tr>
<tr>
<td>0</td>
<td>NOVICE LOW</td>
</tr>
</tbody>
</table>


Thus, where the government scale defines eleven proficiency ranges, the academic scale establishes only nine. The government scale discriminates among four formal higher proficiency ranges than the academic scale which, in turn, includes two additional lower subdivisions.

1.11.15 Language Specific Descriptions

In addition to reformulating the 01 rating scales, ACTFL formulated language specific descriptions for French, German and Spanish expressly for foreign language teachers who want to know what the 01 is intended to measure. For similar reasons, ACTFL also wrote a set of cultural guidelines (Lowe 1984, 34).
1.11.2 Function in the Oral Interview Context

Trained OI testers rate examinees' oral proficiency on the basis of a Functional Trisection of Proficiency Levels, which is subdivided into three aspects: Function, Content/Context and Accuracy. Functions include tasks accomplished by examinees, attitudes expressed and tones conveyed, in other words, what examinees can do with the spoken language.

The Content/Context aspect encompasses topics, subject areas, activities and jobs addressed by examinees. Thus, in the ILR interview, Content is often job specific. Simultaneously, testers must account for Accuracy, the number of errors that examinees make (Heidinger 1983). Yet Accuracy also reflects the acceptability and the quality of the message (Lowe 1984, 36, 42).

Since the OI is intended to measure examinees' functional ability, criteria for sample tasks such as listing, narrating and supporting are indicated in the Functional Trisection. Further, since the OI is an integrative rather than a discrete point test, examinees must combine pronunciation, fluency, vocabulary, grammar, and cultural skill during the interview. Given those constraints, testers must choose stimuli in response to which native speakers would automatically integrate linguistic and cultural skills. It is in such situations...
that testers evaluate examinees' ability to integrate acquired skills into functions.

Ratings at the upper end of the OI scale are non-compensatory. This means that examinees must demonstrate proficiency in all three aspects on the Functional Trisection, which testers use as criteria, in order to be assigned a Superior rating. Only after assigning overall scores based on those criteria may testers examine contributing factors. The factors, in approximate order of significance in the ACTFL Superior range, are pronunciation, fluency, sociolinguistics/culture, vocabulary and grammar. These factor scores are non-cumulative, serving only in an analysis of the overall performance (Lowe 1984, 36-38).

1.12.0 Oral Interview Reliability

Test reliability is "the extent to which individual differences are measured consistently (Sax 1980, 630)." By and large, rater reliability determines the overall reliability of measurements made with the OI.

1.12.1 Rater Reliability

Lowe and Liskin-Gasparro (1983, 2) claim the OI is reliable in that independent ratings of the same live or taped interviews made by trained testers normally differ by no more than a plus point. ACTFL, and presumably the
government too, obtains reliability by monitoring both interrater and intrarater reliability. Having two or more raters independently evaluate the same interview provides a control for the first, interrater reliability. For the second, intrarater reliability, the same rater must rate the same interview on separate occasions. Whenever necessary, ACTFL retrains raters. Lowe (1984, 37) reports that consistent rater training produces a standard error of measurement within a plus point on OI ratings.

1.12.2 ACTFL/ETS workshops

In addition to familiarization workshops from two hours to two days in length, ACTFL and ETS direct tester training workshops from four to five days in length (Lowe and Liskin-Gasparro 1983, 3). To date ACTFL has trained OI testers in French, German, Italian and Spanish, in addition to OI trainers, who lead familiarization workshops in French and Spanish (Lowe 1984, 34-35).

1.13.0 Oral Interview Validity

Several criteria exist to assess the validity of measurement instruments. Whether or not measurements correspond to these various criteria constitutes the instruments' validity.
1.13.1 Criterion Related validity

Criterion-related validity is a general term for two kinds of validity, concurrent and predictive (Sax 1980, 634). For each particular application the OI must be validated according to specific external criteria. Lowe (1984, 37) explains that concurrent validity can be assessed by comparing measurements with classroom and job performance at the time of evaluation. Similarly, predictive validity is the correlation between test scores and certain criteria, but with an intervening time period (Sax 1980, 634).

1.13.20 Construct Validity

The aim of a construct validity study is to determine whether or not measurements support the existence of hypothesized traits. That is, when sets of measurements correlate highly, they measure the same trait. In contrast, when measurement items or instruments do not correlate, they discriminate between individual traits and among irrelevant variables (Sax 1980, 634). Although concurrent validation is common for comparing indirect measures to direct measures such as the OI, Lyle F. Bachman and Adrian S. Palmer (1981, 67-86) assert that adequately defined criteria likely do not exist for comparison with the OI and, therefore, that construct validation is necessary.
Bachman and Palmer used a multi-trait multi-method approach in "The Construct Validation of the FSI Oral Interview". They measured two language traits, speaking and reading, using three methods: interview, translation and self-rating. Their sample consisted of 75 native Mandarin speakers of English. Bachman and Palmer analyzed their measurement data using Campbell-Fiske criteria for convergent and discriminant validity. They found:

The results indicate both convergent and discriminant validity for the FSI interview.... These results provide strong evidence against a model of unitary language competence and support a model of partially divisible language competence."

The findings indicate that the ostensibly measured traits, speaking and reading, converge to a considerable extent. It is for this reason that Bachman and Palmer argue against a completely divisible trait hypothesis.

Confirmatory factor analysis enabled them to determine the effects of trait and method on language proficiency measures. Their data indicated that method had a major influence on the different measures. Yet the OI's speaking trait component was the largest while its method component was the smallest.

In conclusion, Bachman and Palmer state their goals for multi-trait multi-method validity studies such as
this one. They call for further study to define language traits and, for these traits, to develop and perfect measurement instruments, including perhaps self-assessment forms.

1.13.3 Content Validity

The government's OI testing experience has shown that the most variable aspect of the OI is its content. Content may be job specific and may include literary, political, and technical language (Lowe 1984, 36, 42). Nonetheless Lowe and Liskin-Gasparro (1983, 2) contend that testers evaluate functions, content and accuracy at each level as specified in the Functional Trisection and that, in this way, they are able to maintain content validity from interview to interview. Lowe (1984, 36) explains further that testers can control content validity through elicitation which tests language functions in real-life situations. Here content and context seem to overlap.

1.13.4 Face Validity

Face validity is in all likelihood the most subjective validation criteria. It is how relevant and important examinees find test items (Sax 1980, 634). Lowe and Liskin-Gasparro (1982, 2) consider the OI face-valid for speaking ability because, in the OI, examinees must speak in a realistic conversational setting. Face-valid,
then, the OI seems to be a reasonable way to evaluate speaking skill (Lowe 1984, 39).

Thus face validity, indirectly at least, depends upon what examinees expect to face in the Oral Interview. But right now, only the United States government, ACTFL and ETS keep their trainees abreast of what takes place and to whom they administer their oral proficiency metric.

1.14.0 Oral Interview Critique

In the course of the current investigation I have encountered a number of difficulties, drawbacks and limitations regarding the OI. Before going on to examine the next measurement instrument, I would like to review them here in order of increasing importance.

1.14.1 Nomenclature

First it is unclear whether the disparity in recent documents, namely reference to (1) interviewers, (2) raters and (3) testers, is intended by the authors to distinguish among educated language speakers, native or non-native, who have been trained in and for:

1. face-to-face interviewing (and not necessarily rating)
2. rating observed or pre-recorded oral interviews (and not interviewing, per se)
3. planning interviews, writing questions, selecting suitable items and/or examinees, analyzing and reporting test data, and placing applicants or otherwise applying test results (but not necessarily interviewing or rating)

or whether it is imprecision on the part of the authors.
1.14.2 Applications

Second, measurements made with reference to limited content and context objectives, including job-specific government, industrial and academic applications, might not evaluate examinees' overall communicative competence. Even so, Lowe (1984, 38-39) suggests that the OI is most useful for summative proficiency evaluation when, for example, the interviewee has completed a course of study. It may be equally useful measuring proficiency when the interviewee has experienced living abroad.

1.14.3 Minor Drawbacks

The OI, then, would seem to have numerous applications. Lowe (1984, 38), however, recommends against frequently administering the OI to the same individuals. What appears to be the intent of this recommendation is to avoid biased or invalid scores resulting from learned interview behavior where examinees play a knowledgeable role in the measurement process. Those Heidinger (1983) referred to as specials, persons who excel in the interview setting, as others at paper-and-pencil tests, may lack breadth and tend to pose problems for whoever evaluates interviewees' performance.
1.14.4 Effort and Time

Although the OI directly measures speaking proficiency, the ratio of interviewers to interviewees must equal or exceed one-to-one. For this reason the OI, including preparation, rating and review, requires more effort and takes more time than traditional paper-and-pencil tests (Lowe and Liskin-Gasparro 1983, 3). Heidinger (1983) spoke from experience when he reported that the interview exhausts the interviewer as well as interviewee.

1.14.5 Limited Opportunity

Above all, proper use of the OI requires systematic tester training because it is a highly structured procedure. According to Lowe (1984, 37-8) government training lasts two weeks; ACTFL takes four days. The problem is that the trained tester in-group is only now expanding to far away places like Montana. Professional groups such as the Pacific Northwest Council on Foreign Languages (PNCFL 1985, 1; Gonzalez 1984) have been clamoring for information on and training in OI measurement.

Nonetheless responsibility for measurement in foreign language classes belongs to individual language instructors. It is our responsibility to acquire basic knowledge of and a more widely practiced proficiency in oral testing. Depending on interest, opportunity and the extent of training and practice, language instructors as
well as language learners should become involved in the OI, for as Lowe and Liskin-Gasparro (1983, 3) conclude, "the importance of oral competence fully justifies the time and effort required to test it."

1.20 The Test of Spoken English

ETS offers an alternative to the Oral Interview to institutions wishing to evaluate English speaking proficiency. This alternative, the TSE, is a group test administered under standard conditions and scored by ETS. Below I describe and evaluate the TSE, relying heavily upon ETS for the most up-to-date information on this, ETS's proprietary instrument.12

1.21.0 General Description

The standardized administration of the TSE has circumvented face-to-face interviewing by eliminating direct conversation between interviewees and trained native or near native speakers acting as interviewers. Yet ETS (TSE 1982, 5-6) asserts:

The format of the TSE retains the high degree of validity inherent in the direct interview procedure while virtually eliminating the subjective measurement problems associated with interviewing.

In spite of the standardization of the TSE, ETS aims to test "linguistic content . . . appropriate for all examinees regardless of their native language or culture."13

27
While trying "to elicit evidence of general speaking proficiency", ETS (TSE 1982: 5, 7) claims to have avoided questions hinging upon "extensive familiarity with a particular subject matter or employment context," a criticism leveled against the Oral Interview above. ETS claims that the TSE is not similarly targeted, "that it permits the examinee, regardless of native language, to demonstrate general speaking proficiency," and, further, that TSE subscores may be used diagnostically.

1.21.1 TSE Development

ETS developed the TSE chiefly for use by academic institutions, the only context where it has been validated. Additional validation research was being pursued, as of the 1982 date of publication, in health-related professions.\(^1^4\)

In an attempt to support a basic premise of its standard test development, ETS (TSE 1982, 5-6) cites a 1979 study of an outdated TOEFL form to substantiate:

> correlations between paper-and-pencil tests of listening comprehension, reading, and writing skills and tests directly measuring active speaking ability . . . on a group basis . . . . However, extrapolating on the basis of group correlation data to statements about the speaking proficiency of individual test candidates is questionable.

So ETS recognized the potential disparity between an individual's TOEFL score and the individual's ability to communicate in spoken English before it developed the TSE as an adjunct to its TOEFL program.
ETS constructed a prototype test of spoken English that included items measuring discrete language elements, in English lexicon, morphology, and syntax, as well as free response items:

The specific items within each item type were selected with the goal of maintaining the highest possible correlation with [Foreign Service Interview] rating and lowest possible correlation with TOEFL score.

ETS (TSE 1982, 7-8) sought items with the least correlation to Foreign Service Interview grammar and vocabulary sub-ratings, and to "sections 2 and 3 of TOEFL which involve written language only."15

ETS (TSE 1982, 19) documented only a small sample (N=60) which it claims depicts the relationship between scores on the Foreign Service Interview oral proficiency interview and performance on the TSE. In that study, the approximate relationship was tabulated on the basis of pre-1980 data.14 By the end of 1985, however, roughly two thousand individuals will have taken the TSE (Ballard 1985).

1.21.20 TSE Format

The TSE, described by ETS (TSE 1982, 8-10) in 1982, still consists of seven sections:

1. unscored warm-up
2. reading aloud
3. completing sentences
4. story-telling from drawings
5. answering questions about drawing content
6. responding at length:
   a. describing common objects
   b. expressing opinion
7. describing an imaginary course schedule

1.21.21 TSE Mechanics

The standard administration of the TSE requires tape playback and recording equipment. According to ETS,

The test takes approximately 20 minutes and can be administered to individuals with cassette or reel-to-reel recorders or to a group using a multiple-recording facility such as a language laboratory.

A test book, which complements the test tape and the individual response tapes, contains general directions for the examinee to read. In addition to the general directions, each section of the test has special instructions, which I presume must also be read by the examinee. Although it is difficult to imagine an examinee unable to read English who takes this test, reading comprehension, that is, not understanding what to do, definitely could influence examinees' performance adversely. Their performance, during a standard administration at least, is recorded non-stop on individual response tapes.16

1.21.22.0 TSE Records and Reports

ETS compiles and then reports a considerable amount of data on individuals and examination results. This information has a number of forms and uses.
1.21.22.1 TSE Scores

First of all, ETS collects and distributes individuals' test scores:

Official score reports include . . . four different test scores: a score for overall comprehensibility and scores for three diagnostic areas—pronunciation, grammar, and fluency. Since each score is derived from a different rating scale, diagnostic scores do not contribute to the overall comprehensibility score.

ETS bases these four scores on four separate criteria (TSE 1982, 11). Though ETS records and stores TSE scores, they can be individually identified only for a year and one half. During that period, ETS reports scores to institutions at the request of examinees. After that, "TSE score data that may be used at any time (emphasis ETS's) for research or statistical purposes are not individually identifiable (TSE 1983, 12)."

1.21.22.2 Other Reported Data

In addition to the four scores mentioned above, ETS records and reports, among other items, the native country and language of examinees as well as their date of birth (TSE 1982, 11). ETS uses a numeric code to distinguish among six geographic regions of language origin indicated by the first (left-most) digit as follows:

100: Africa
200: Americas [native languages]
300: Asia [including Japan]
400: Europe
500: Middle East
600: Pacific
With certain exceptions (176--Kiswahili, 203--Arymara, and 436--Georgian), second order numeric coding (second and third digits from the left) follows alphabetical order by region (TSE 1983, 14).19

1.21.22.3 TSE Norms

In its 1982 Manual for Score Users (TSE 1982, 13-14), ETS reported the score norms for a population of 610 examinees who took the TSE from October 1979 through March 1982. The sample population included sixty prospective veterinarians and 550 teaching assistants or applicants for teaching assistantships. Score comparisons for overall comprehensibility and for each of the diagnostic areas were published with the following caveat: "it cannot be assumed that the [score comparison] table accurately characterizes the relative language proficiency of all teaching assistants." In addition, there appears to be an error in the mean reported for Grammar in the table on page 14 (TSE 1982).

The TSE Examinees Handbook (1983, 11), however, does state:

At the time you [the examinee] receive your score report, you will receive additional information that will help you interpret the scores.

I asked Rod Ballard, TSE Program Director, if that meant that norms for language groups and for country of origin (such as those for the TOEFL) had recently been made avail-
able. No, Ballard (1985) said, examinees score reports simply listed the rating scales on the back of the documents. Since, in the TSE's short history, there had been fewer than 2000 applicants, complete statistical norms were not yet available. This is unfortunate because group norms, for examinees' and institutions' reference, constitute a major advantage of standardized tests.

1.21.30 TSE Function

ETS intended the TSE to assess examinees proficiency at moderately complex speaking tasks, that is, proficiency which is roughly equivalent to Oral Interview Levels 2 and 3. Now, because it is difficult to compare these two instruments critically without an intimate knowledge of both, a closer look at the seven sections of the TSE is in order. Here, and again in the following sections, I rely upon ETS documentation (TSE 1982, 8-18).

1.21.31 TSE Section One

For no score, in Section One, examinees respond to taped biographical questions. For rating purposes, however, TSE Section One could serve to inform the rater about examinees' ability to conform responses to the standard administration format.
1.21.32 TSE Section Two

In Section Two, examinees read a written passage aloud, after preliminary silent reading. ETS raters score examinees' taped production for "pronunciation and overall clarity of speech".

1.21.33 TSE Section Three

Examinees' performance in Section Three consists of sentence completion which must be meaningful and grammatically correct. It is not clear from ETS's description whose meaning examinees must convey nor what standard ETS uses for "grammatically correct" utterances.

1.21.34 TSE Section Four

Story-telling constitutes the performance that ETS requires in Section Four. From a series of six illustrations approximately one and one-half to two inches by two and one-half inches in size, examinees are "asked to tell the [emphasis mine] story that the pictures show, and to include as much detail as possible."22

1.21.35 TSE Section Five

In Section Five examinees answers successive questions about the content of a single illustration.22 The questions are prerecorded on the test tape.
TSE Section Six

Section Six includes open-ended items. In this section examinees give longer responses, both describing common objects and expressing opinions on issues with which they should be familiar.

TSE Section Seven

Again in Section Seven, description is the performance that ETS requires. Examinees must produce a spoken description of a printed schedule which appears in the test book, as if explaining it to a class at the first class meeting. Thus, in both Section 3 and Section 7, ETS requires spoken responses to written stimuli.12

TSE Reliability

One problem with evaluating the TSE or other similarly controlled proprietary measurement instruments lies in the fact that most reliability and validity data come from internal studies. By that I mean that the data are reported by the organizations, and perhaps even the persons, that developed the instruments themselves, in this case, ETS, and which market their monopolistic services to the academic community. Test taking and test results do not come free, nor do ETS research reports, these days. Internal data and studies scream for independent verification and confirmation with reference to current testing instruments. Although reliability, as well as validity,
data in the 1982 TSE Manual for Score Users predate the five year period set forth in the introduction to this chapter for consideration here, I have decided, in lieu of more recent data, superficially to reproduce the ETS discussion relevant to interrater reliability.

1.22.1 Rater Reliability

Rater reliability is a measure of the effective training of raters. ETS trains official TSE raters in one-day workshops, and then re-trains experienced raters annually, "or more frequently if score discrepancies indicate" that retraining is required.

Two ETS raters make independent ratings of each examination tape. If the two differ substantially in rating one or more diagnostic areas, ETS has a third rater listen to the tape. However the third person rates only those areas on which the two initial raters differ. In the case of complete disagreement among the three, the two highest scores on the disputed diagnostic areas will be averaged for reporting purposes (TSE 1982, 10). Further, retraining one or more of the raters might be in order.

1.22.20 "Internal Consistency"

The other reliability judged relevant by ETS was the "internal consistency" of reported scores.
1.22.21 Pre-1980 Reliability

ETS used Cronbach's alpha coefficient methodology to measure "an internal consistency reliability of .91 and a standard error of measurement of 15 (Clark and Swinton 1979, 53)." Given the simple fact that the standard error is nondecimal, and greater then three, this standard error must apply to the overall comprehensibility score, which is reported an a different scale than the diagnostic scores.

1.22.22 Reliabilities and Standard Errors

ETS (TSE 1982, 18-21) omits mention of internal alpha coefficients and standard errors for the three TSE diagnostic area scores. Additional information from ETS on these points could shed light on the question whether, by any standard statistical method, those diagnostic scores can be shown to reflect individuals' true ability to perform against consistently measurable criteria.

1.23.0 TSE Validity

Test validity studies indicate how accurately tests actually measure what their authors intend them to measure. The drawback is that there is neither a single indicator nor a standard index of test validity. ETS (TSE 1982, 21-22) refers to three types of validity which it judges relevant to study of the TSE: content validity, construct validity and criterion-referenced validity.
However, ETS circumvents critical discussion of content validity by referring the reader to TSE Content and Program Format (TSE 1982, 8-16). Likewise, those interested in construct validity ETS directs to Development of TSE (TSE 1982, 6-8). Once again, the information which ETS (TSE 1982, 21-22) presents predates the chronological range of this study. Nonetheless, it must be considered in lieu of more recent data and independent findings.

1.23.1 Construct Validity

In an attempt to avoid positive correlations between the TSE and the TOEFL (ETS's proprietary instrument), and between the TSE and the Oral Interview grammar and vocabulary subsections, ETS (TSE 1982, 8-12) constructed a different grammar metric, the TSE grammar scale:

Not surprisingly, the TSE grammar score is more closely associated with the TOEFL total than are the other three TSE scores. This suggests that the TSE grammar score is tapping an ability similar to that measured by the TOEFL, which is a paper and pencil multiple choice test (TSE 1982, 23)."

This shows that results diverged markedly from a stated aim of TSE development; ETS wanted to decrease the correlation between the TSE and its written partner, the TOEFL.

There is an undesired correlation between the TSE overall comprehensibility score and the TOEFL Total score calculated by ETS at .57. This falls in the same range as the desired correlations, from .52 to .60, with formally assessed "ability to handle common situations involving
language skills." Similarly, ETS (TSE 1982, 23-24) reported that the TSE "overall comprehensibility rating is more closely related to pronunciation and fluency . . . than to grammar" and claimed that "while TSE scores are closely related, each provides some independent information about examinee's proficiency." But the intercorrelations, undesirable for discriminant validity, range higher (.79-.93) than the criterion-referenced correlations with Oral Interview ratings (.73-.79), which I discuss further in subsequent sections.

Even though TSE grammar score correlations to the other TSE scores ranged from .79 to .84, ETS suggested that they diverged sufficiently from one another for diagnostic purposes, with the scores thus taken to represent differentially measurable skills. Differentiation of skills is an aspect of construct validity. On the other hand, ETS contended that correlations ranging from .52 to .68 validated the TSE in instructional settings relative to the GTA's "ability to handle common situations involving language skills", that is, where convergence is to be desired. The answer to the question how, while higher statistical correlations in the above instances affirm divergent validity between internal constructs, lower statistical correlations confirm convergent validity simply escapes me.

It appears as if ETS judged the validity of their own
instrument according to various criteria (ranges of correlations) in the context of desired versus undesired correspondences with external and internal constructs.

1.23.20 Criterion Related Validity

A certain amount of overlap is to be expected where correlations support, respectively, construct and criterion-related validity claims. ETS (TSE 1982, 23-24), in the following study, seeks to correlate TSE ratings to external criteria, to Oral Interview ratings.

1.23.21 A Validity Study

For an extremely limited sample population (N=31, before 1980), the TSE pronunciation score showed the highest correlation with the FSI oral proficiency rating. That correlation was .77. At the same time, however, one should consider the .79 intercorrelation between TSE pronunciation and TSE grammar, which had the highest correlation to TOEFL Total Score (.70). There are two possible interpretations, either that TSE scores actually approximated Oral Interview measurements or that the government testing program measured the same skill as ETS's TOEFL Program, written English grammar and vocabulary. It is of interest to note, in this context, that the government program has been adapted and revised since then.
1.23.3 Three Observations

In closing this description and evaluation of TSE reliability and validity, I would like to make three observations. First, ETS's reliability and validity data are not current; second, they appear insignificant in sample size; and, third, they have been interpreted in the most favorable light by ETS. These observations lead into the next section, a critique of the TSE.

1.24.0 TSE Critique

In this review of ETS documents on its own instrument, the TSE, I take a grain-of-salt approach because ETS developed it a good five years ago. Since that time I have observed little change in the TSE, by outward appearances.

1.24.1 Reading Skill

ETS (TSE 1983, 7) provides examinees more detailed written instructions in Sample Questions than actually appear in the testbook. Even so, examinees must read to take the TSE under standard administration. "Thus, for a blind examinee no diagnostic score can be reported for pronunciation or for fluency (TSE 1982, 16)." This examination of spoken English, then, does not appear face-valid when the two scores valued for speaking by ETS depend so heavily upon reading skill.
1.24.2 Illustrations

Illustrations do appear in the TSE Examinee Handbook. These photographic illustrations are so small as to require close scrutiny for detail.22

1.24.3 Time Limits

Time limits on each section could influence scores by restricting examinees' response times. ETS (TSE 1982, 19) asserts that speed is not a factor in completing the formally administered TSE, but time, which probably serves as a variable in evaluating Fluency, is controlled by the playback and recording speeds during administration.16 Although a "voice on the test tape announces the beginning and the end of each section", examinees are not necessarily advised in advance of the time allowed to respond to each question in Sections One, Five and Six (TSE 1983, 7-10). Quite conceivably, time limits could contribute to a ceiling effect mentioned by ETS (TSE 1982, 19).

ETS (TSE 1983, 8-10) specifies silent and oral reading times of one minute each for TSE Section Two, as well as picture study times of one minute each for Sections Four and Five, and silent schedule reading and oral reporting times of one minute each for Section Seven. Thus ability to plan, speak and time oneself simultaneously could yield examinees a distinct advantage in addition to reading ability. It is inconceivable that raters would
score responses continued during subsequent questions. It
is equally inconceivable that examinees would not shut up
and listen to the voice on the test tape. It is easy to
see that administration speed is controlled by ETS design.

1.24.4 Targeting

Examinees' performance in TSE Section Seven mimics
that in a narrowly targeted scholastic setting with a re­
stricted audience. The role that this item plays in a test
of general speaking proficiency is unsupported by ETS's
validation study undertaken in English-Medium Instructional
Settings. For a direct measure of communicative com­
petence, Section Seven suffers from direct targeting of a
particular academic milieu. The weights of individual test
items in the diagnostic scores and in the overall compre­
hensibility scores could easily have been juggled during
development to produce the desired statistical correlations
in this particular setting. How well scores correlate with
external criteria for larger samples in experimentally
varied communicative settings remains anyone's guess.

1.24.5 Content/Context

Test items fix the content of examinees' responses
in Sections Two, Four, Five and Seven. Further, context is
determined by items in Sections One, Two, Three, and Seven.
The only open-ended questions appear to be in Section Six,
where ETS does not score the actual ideas. Even in this
section, however, ETS limits content by reference to common articles and familiar topics (TSE 1982, 8-9; 1983, 8-10). But telephone sets and automobile pollution might not exist in all cultures and societies.\textsuperscript{13}

1.24.6 ETS Caveats

Under its caveat, "Do not use TSE scores as the sole indicator of potential teaching performance [ETS emphasis] (TSE 1982, 15)"., ETS distinguishes (English) language proficiency from interpersonal skills and interest, social factors and individual traits not measured by the TSE. ETS (TSE 1982, 24) also warns those who make decisions affecting individuals based on TSE scores that they should establish validity in situ, for their own specific purposes, beforehand. This holds true for all applications of standard measurements.

1.24.7 Native Language/Country

If the TSE were validated in a wider range of communicative settings than college classrooms, and if the sample examinee population were broadened to include more than GTAs and the like, norms based upon native language group performance could provide helpful insights to English-as-a-Second-Language instructors, programmers, and writers. With error analysis they could pinpoint acquisition difficulties for particular groups of language
learners identified by native language and country. A dangerous potential, yet, may lie in ETS reporting linguistic and geographic origin data with individual test scores to institutions, for the data could be used for discriminatory practices as well as for statistical purposes.

1.24.8 TSE Currency

A discrepancy probably exists between the TSE forms used for the purpose of gathering reliability and validity data prior to 1980, and the TSE form in current use, since ETS markets forms that it no longer uses in the TOEFL program. Those forms are sold to institutions, packaged in the Speaking Proficiency English Assessment Kit (SPEAK).

The SPEAK sells for three hundred dollars and contains in-house rater training and practice materials (including tapes) in addition to test materials. A second dated test form is available to institutions that have already purchased the three-hundred-dollar kit, for an additional seventy-five dollars. 27

1.24.9 Critical Limitations

The TSE shows limited potential for applications at elementary proficiency levels as well as at the advanced level because,
the TSE is pitched at a level beyond that of mere survival skills, but does not require the breadth of vocabulary and general appropriateness [emphasis mine] of idiom expected of the well-educated native speaker.

Though the TSE adequately may measure differences between examinees whose Oral Interview ratings would be between two and three plus, it serves its function less well when distinguishing among advanced English speakers because above government level three plus a ceiling effect begins to appear (TSE 1982, 19).

The intermediate difficulty of the TSE closely approximates, in English proficiency, what Heidinger (1983) said was to be expected of college graduates with foreign language majors, Oral Interview ratings between two and three plus. But even at this level, test users should avoid rigid cut-off scores due to imperfection inherent to all measurement instruments. The standard error reported on earlier TSE forms, and presumably obtained with experienced raters, showed considerable variability in individual measurement (TSE 1982: 15, 19, 21).

1.30 The Advanced Placement Examination in French Language (APFL)

The APFL form described in the May 1985 Advanced Placement Course Description (CEEB 1984) remains essentially unchanged from that described in the 1984 edition. The latter publication dates from 1983. While the comments and criticisms that I make pertain primarily to the APFL
Speaking Subsection, I wish to underline the fact that the College Board has reported no substantive changes in this ETS proprietary measurement instrument since 1983.28

1.31.0 APFL Form

The APFL is written in French with the notable exception of the directions (CEEB 1984, 3). The College Board contends that the directions must be in English since the College Board uses this examination to measure American high school students' proficiency in the French language, and English directions probably do facilitate administering the test to native English speakers. In spite of this discrepancy (English in a French proficiency measurement instrument), examinees must respond exclusively in French.

1.31.1 APFL Speaking Subsection Format

The Speaking Subsection measures French speaking proficiency in much the same way that the TSE measures English. Test administrators play a cassette tape recording that provides stimuli (and times the subsection), while examinees continuously record their spoken responses. The College Board advises instructors, administrators and examinees alike that unfamiliarity with recording equipment could cause administrative problems during the examination.

Although the Speaking Subsection takes only twenty minutes to administer during a two and one-half hour examination, examinees' speaking proficiency makes up one
quarter of the APFL total score (CEEB 1984, 3-4). This is disproportionately high if you consider the test time devoted to the measurement of speaking skill.

1.31.20 APFL Speaking Subsection Function

The purpose of the Speaking Subsection is to measure examinees' ability to respond correctly and idiomatically in French (CEEB 1984, 3).

1.31.21 Directed Responses

First examinees must respond to questions and directions. The test tape allows examinees only fifteen seconds to respond to individual items. The College Board then employs scorers who evaluate such short answers against the following criteria: appropriateness, grammatical correctness and pronunciation (CEEB 1984, 14).

1.31.22.0 Story Telling

Subsequent to making directed responses, examinees must recount a story in response to stimuli which consist of line drawings. This item shows remarkable similarity to TSE Section Four. Examinees may study the drawings for two minutes before recounting, in two additional minutes, the complete story illustrated by the drawings.
1.31.22.1 Illustration

The same sample drawings appear in the 1985 description (CEEB 1984, 16-17) as in the 1984 description (CEEB 1983). ETS, the copyright owner, provided them from an even earlier examination form. These line drawings are approximately two inches by three and three-quarters inches in size. The drawings include details, potentially crucial to the story, a lost pet for instance, that measured barely one-quarter inch in size. If it were just a story examinees were to tell, such detailed illustration content might not figure in the scoring.

1.31.22.2 Scoring Criteria

It is difficult to ascertain, due to slight differences in wording, whether the scoring criteria for story-telling expand upon or differ from those for directed responses. Raters score the story which examinees recount "not only for appropriateness of vocabulary and grammatical correctness but also for the fluency, range of vocabulary, and pronunciation . . ." evident from examinees' recorded responses (CEEB 1984, 14-17).

1.32.0 APFL Reliability

The reliability of the Advanced Placement Examinations, in general, depends heavily upon the security of the multiple-choice questions used from year to year. This College Board admission immediately calls the
reliability of the free-response sections, including the APFL Speaking Subsection, into question.

1.32.1 Free Response Scoring

After the examination, high school and college language instructors meet to score examinees' response tapes. General criteria for scoring free-response sections are initially developed by group leaders employed by the College Board from year to year. Scoring group consensus then determines detailed criteria or, at least, expectations for a given test administration.

Raters assign scores independently, while group leaders continuously monitor their work. This constant monitoring probably produces a respectable degree of inter-rater (intra-group) reliability for scores on a given free-response section in a given year.

1.32.2 Final Grade

Finally raters combine multiple-choice and free-response scores to arrive at a final score which the College Board reports on a five-point scale:

5—Extremely well qualified
4—Well qualified
3—Qualified
2—Possibly qualified
1—No recommendation
Colleges and universities may then use these recommendations for advanced placement purposes according to locally determined guidelines (CEEB 1984, iii-iv).

1.33.0 APFL Validity

The College Board offers reprints of an APFL study to those educators interested in the French language examination's validity (CEEB 1984, 2). Unfortunately this study, published in 1975, predates numerous changes in the instrument, its format, and in the functions tested.  

1.34.0 APFL Critique

The overall score for this particular Advanced Placement Examination combines scores for four skills: listening and reading, tested with multiple-choice items, and writing and speaking, tested with free-response items. Each skill weighs equally (CEEB 1984, 3). Thus one-half of the final grade depends upon free-response scores resulting from evaluation against consensus criteria. With the more reliable multiple-choice scores accounting for only one-half of the final grade, the reliability of the combined test results remains open to speculation.

1.34.1 Consanguinity: Speaking Cousins

Although the Advanced Placement Program is the work of the College Board, the APFL belongs to ETS. The similarities between the Speaking Subsection and the TSE,
another ETS instrument, are striking. There are also
differences. For one, ETS dropped oral reading from the
APFL Speaking Subsection some time ago.30

While data are lacking on the reliability and the
validity of this subsection alone, ETS did not design it to
stand alone. Yet the APFL Speaking Subsection alone
appears to depend less upon reading skill than the entire
TSE, which ETS designed to measure speaking proficiency
alone.

1.34.2 Scoring

The Advanced Placement Course Description: French
booklet, and perhaps the APFL Speaking Subsection as well,
notably lacks explicitly stated criteria for evaluating the
free-response section of the APFL, and for the Speaking
Subsection in particular. From what the College Board
reported about its free response item scoring procedure, I
gather that criteria vary from year to year according to
scoring group consensus (CEEB 1984, iii). For five
dollars, though, the Advanced Placement Program does offer
a supplemental booklet entitled Grading the AP Examination
in French Language. Investment in that booklet could yield
explicit criteria to scrutinize more closely.31
CHAPTER ONE

NOTES

Section 1.0


Section 1.1

3See Section 1.00 for a brief introduction to the DLI or government scale. In the most recent treatment studied here, Pardee Lowe, Jr. (1984) refers to the ILR rating scales which I assume differ little, if at all, from the DLI scales (Lowe and Liskin-Gasparro 1983).

4See APPENDIX A for complete definitions of each level on the Government (DLI) Rating Scale (from Liskin-Gasparro's 1983 compilation).
See Lowe (1984, 32-34) for a synopsis of OI development; see APPENDIX B for ACTFL/ETS Provisional Speaking Descriptions (from ACTFL's 1982 compilation).

See APPENDIX C for the ILR Functional Trisection, ACTFL Provisional Guidelines: Provisional Generic Descriptions and Provisional German Descriptions (from Lowe 1984, 42-49).

See Section 1.14.2 for a brief discussion of particular applications; see CHAPTER 3 for further discussion of OI applications.


See APPENDIX C for the Functional Trisection; see also Section 1.11.2 for discussion of function and context and Section 2.22.21 for discussion of setting in the OI.

See Sections 2.20 through 2.23 for further discussion of the relationship between the OI and communicative competence testing.

See CHAPTER 3 for further discussion of OI applications.

Section 1.2

In April 1985, Educational Testing Service (ETS) provided the 1982 edition of the Test of Spoken English (TSE) Manual for Score Users (Princeton, N.J.: ETS, 1982), to Dr. Robert Hausmann, the ESL Program Supervisor at the University of Montana; Rod Ballard, TSE Program director, reported in a telephone interview May 1985. that a revision was planned for 1986; see Section 0.21, in INTRODUCTION.

The question whether standardized testing successfully avoids linguistic or socio-cultural bias lies beyond the scope of the present study.

See Sections 1.23.0 ff., for further discussion of TSE Validity.
See TSE 1982, 6-8, for details of TSE development; incidentally, the criteria based upon TOEFL sections inform critical limitations of ETS's own TOEFL treated in an unpublished paper by Beaufait (1984).

See TSE 1982, 8-10, for detailed ETS explanation of content and administration, including non-standard administration for handicapped examinees.

This aspect of TSE scores mirrors the non-additive interrelationship found between Oral Interview factor scores and the overall oral proficiency rating (see Section 1.11.2 and Section 2.13); see APPENDIX D for TSE rating scales.

Such data could conceivably serve parties (internal or external to ETS) in discrimination by age or origin.

These exceptions probably originated from previously uncoded native languages reported by recent applicants to the TOEFL or TSE testing programs.

There are two possible further applications of language/country of origin norms beyond information for institutions and examinees: (1) monitoring linguistic and/or cultural discrimination by test items or the instrument as a whole, and (2) diagnosing English acquisition difficulties common to homogeneous populations, i.e., adult native Japanese speakers of English as a Second or Foreign Language.

See APPENDIX A for the most recent government descriptions of Oral Interview levels; see also Sections 1.11.11.0 and 1.11.11.1, for discussion of government levels.

There is a discrepancy between the description of TSE Section 4 in the TSE Manual for Score Users (1982, 8) and that in the TSE Examinee's Handbook (1983, 7-8): the first specifies "line drawings", while the second provides the examinee sample photographic illustrations; this discrepancy appears again between descriptions of TSE Section 5, respectively; see Sections 1.21.34-35 and 1.24.2, in the present study.

See Sections 1.24.0 ff., for a critique of TSE functions.

'Raters' are also referred to as 'testers' by Lowe (1984) and others; see Section 1.14.1.
Since rating scales differ dramatically, only a rough comparison can be drawn to the one 'plus' point standard error reported for the OI (see Sections 1.12.0 ff.); this approximation is supported by ETS's own comparison (TSE 1982, 18-19).

John L. D. Clark and Spencer S. Swinton, 1980, The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings, TOEFL Research Report 7 (Princeton, NJ: ETS); Clark and Swinton were also active in the development of the TSE; see Section 1.23.21.

Rod Ballard, TSE and SPEAK Project Director, telephone conversation 26 April 1985; the current TSE fee for standard administration is $40 (1984-1985 Application Form).

Section 1.3

The Advanced Placement Program quit offering reel-to-reel student response tapes as of May 1984 (CEEB 1984, 3); the program does offer item number 464588, Tape recording (French Language, form E, 1982), price: $6; see Note 31 for address.

This part of the APFL Speaking Subsection differs from TSE Section 4 in that the sample illustrations are still line drawings; the picture study time and storytelling time of two minutes each compare to one minute each for the TSE; see Sections 1.21.34 and 1.24.3.


Item number 235487, Grading the AP Examination in French Language, price: $5, is available from the Advanced Placement Program, Box 2899, Princeton, NJ 08541.
Among viable hypotheses advanced for language competence, there are two which exclude the possibility of each other. These are that competence is either indivisible or that it is partially divisible. Conclusions generally favor the hypothesis for indivisible language competence (Scholz and others, in Oller and Perkins 1980, 24-33). What guides the discussion in this chapter is how hypotheses about the nature of language and how testing strategies figure prominently in oral proficiency as well as in communicative competence evaluation. The progression in this chapter follows testing developments from oral skill, strictly speaking, to communicative competence, in its numerous parameters.

2.10 Oral Proficiency Measurement

By examining a number of hypotheses, studies and guides in this first section on oral proficiency, I intend to prepare ground for more far-reaching discussion of communicative competence evaluation, especially of developments that have taken and are taking place abroad.
2.11.0 An Intensive English Language Program Study

At the University of Illinois, investigators studied an intensive English language program's applications of oral proficiency testing (Hendricks and others, in Oller and Perkins 1980, 77-90). They compared the results of a Foreign Service Institute variant interview technique with three instruments that demanded less time and effort to administer.

Their single versus multi-factor analysis favored a unitary language competence model. Linguistic aspects measured separately, and oral proficiency in particular, could not be clearly distinguished on the basis of test results. The authors suggested, however, that language testing departures from discrete-point measurement will produce feasible alternatives to the Oral Interview.

2.11.1 Discrete-Point vs. Cloze Testing

Hendricks and others at the University of Illinois concluded that discrete-point tests lack detailed socio-linguistic and linguistic context because discrete-point items inadequately assess a person's functional language ability in a realistic setting. Hendricks and others also believe that written and oral cloze tests show promise for refinement to measure parts of what the Oral Interview now does.²
2.12.0 French, Spanish, and Russian Proficiency Measurements

Although specific to only a few languages, two guides to the Oral Interview, one United States Foreign Service and the other ACTFL affiliated, are available.

2.12.1 French and Spanish Guide

The Foreign Service has published a guide for French and Spanish language proficiency interviewers (Adams and Frith 1980). This guide includes comments and suggestions on testing procedures, suggested topics for Oral Interviews and hints for evaluating lower speaking proficiency levels in French and Spanish. These hints, in particular, should interest teachers who need to evaluate beginning foreign language speakers.

2.12.2 Russian Proficiency Handbook

Oral Proficiency Testing in Russian (Wing and Mayewski 1983) is a handbook for a one-day workshop. It provides prospective testers an overview of oral proficiency assessment principles and techniques. The workshop focuses attention on tester training opportunities in Russian, on culture's role in the interview and on Russian specific language situations. The handbook includes ACTFL/ETS question types and provisional generic proficiency guidelines, which, according to Liskin-Gasparro, ACTFL is currently revising.
2.13 Indivisible Competence

Within the relatively limited scope of oral proficiency measurement, a rater-reliability study supported the indivisible competence hypothesis. Karen A. Mullen's study (in Oiler and Perkins 1980, 91) showed that each Oral Interview subscale (Listening, Pronunciation, Fluency, and Grammar) contributed significantly to the Overall score. Thus the composite score represents the best metric for oral proficiency.

2.14.0 Native Speaker Studies

Native English speakers figured prominently in the following two studies. While in the first they were evaluators, in the second they were evaluated.

2.14.1 An Accented English Proficiency Study

In the first study, oral proficiency in English as a Second Language (ESL) was evaluated in terms of non-native accent. When Donn R. Callaway (in Oiler and Perkins 1980, 102) compared naive native language evaluators and English-as-a-Second-Language instructors, he found that they were equally reliable in their judgments of proficiency demonstrated by speakers whose English was variously accented.
2.14.2 Native Speaker Error

Michelle Fishman's conclusion (in Oller and Perkins 1980, 187) was that when pushed to the limits of their ability, native English speakers made the same kinds of errors (in taking dictation) that non-native speakers make. This may help explain Callaway's findings.

2.15.0 Success Factors

In an effort to identify factors contributing to success in language acquisition, Sadako O. Clarke (in Oller and Perkins 1980, 217) found that Modern Language Aptitude Test (MLAT) results did not consistently predict achievement in spite of correlations with other measures of achievement. It should become clear, through further studies, that although the predictive validity of the MLAT was not supported by this study, personal factors do contribute to language proficiency.

2.15.1 Behavioral and Attitudinal Correlates

When Mitsuhisa Murakami (in Oller and Perkins 1980, 227-30) compared native Japanese behavior and attitude correlates with English listening and reading proficiency demonstrated on cloze tests, he made no study of behavioral and attitudinal correlations with speaking proficiency. Yet his study showed that Japanese students demonstrated remarkable progress in listening comprehension.
in the United States, which perhaps only reflects a bias towards reading in Japanese English education.

2.15.2 Self-Reported Data

John W. Oller, Jr., Kyle Perkins and Mitsuhisa Murakami (in Oller and Perkins 1980, 233-40) studied self-reported data including reasons for studying English. They evaluated learners' motivation for English language study in terms of integrative (understanding Americans for example) versus instrumental (getting a good job) motives. Moderate to low correlations among self-reported data, dictation and cloze measures suggested that the reporting questionnaire was in itself a language test. Further, Oller, Perkins and Mitsuhisa found that questionnaire techniques, with respect to language (native or target) and format (oral or written) lack construct validity.

2.15.3 Integrative vs. Instrumental Motives

In yet another study of factors contributing to oral proficiency, Thomas Ray Johnson and Kathy Krug (in Oller and Perkins 1980, 241) measured integrative/instrumental motivation with a Gardner and Lambert instrument and with a redundancy index based on an integrative Foreign Service Institute type Oral Interview. They observed weak to contrary correlations among these measurements and attained proficiency demonstrated in Oral Interviews. They had to conclude, nonetheless, "that affective factors are
important regardless of the difficulty of measuring them," because standard theories could not explain their data.

2.20 Toward Communicative Competence

What follows reflects evaluation of the Oral Interview in terms of its ability to measure communicative competence. The reporting reflects evaluation made by investigators intimately knowledgeable of Oral Interview procedures.

2.21.0 Negotiating Understanding

In English-as-a-Second-Language (ESL) oral proficiency interviews, examinees and testers can negotiate understanding. Since the interview procedure allows testers to evaluate examinees' participation in oral interactions, the Oral Interview could be an excellent place to evaluate more than examinees' ability to understand, that is, to evaluate examinees' communicative competence (Sanders 1981).

2.21.1 Communicative Features

Sara Lyles Sanders realized that even students lacking in oral proficiency could use strategies to negotiate understanding on both sides, theirs and testers, in the Oral Interview. As the basis for her work, Sanders used conversation analysis to examine three features of communicative competence:
1. ability to respond to different question types
2. ability to produce expansions and accounts
3. ability to perceive and focus topic

Her analysis thus evaluated features pertinent to Oral Interview ratings.

2.21.2 Implications

Sander's proposal for rating Oral Interviews would allow even examinees who produce ungrammatical utterances to receive high interview ratings. She suggests:

a totally different approach which is specifically related to examining the student's ability to communicate. The goal is to look for successful interaction even when it is achieved with speech marked by poor pronunciation, limited grammatical structure, and numerous false starts.

Sanders concluded that speakers of ESL have no need to sound like native speakers: "Their greatest need is to be able to communicate in English with whatever control over the language they have." Thus testers should neither employ nor deploy the oral interview to evaluate correct utterances compared with those of educated native speakers. Testers' intent, then, would be to recognize effective communicators, not proficient test takers or educated native speakers.

2.22.0 Sociolinguistic, Interpersonal and Pragmatic Variables

Sociolinguistic, interpersonal and pragmatic variables presuppose linguistic and paralinguistic
features. Of these variables relevant to communicative competence testing, three are crucial: (1) setting, (2) interlocutors and (3) topical and effective outcome (Clark 1982).

2.22.1 Setting

According to Clark, the range of communicative settings encompassed by the Oral Interview can be broadened by testers' use of relevant props. Such props, graphic or figurative, could play an important part in communication between examinees and testers.

2.22.2 Interlocutors

Clark also suggests that testers incorporate role playing exercises to the Oral Interview. Role play can be used to simulate a wider variety of interlocutors, thus enhancing the range of communicative settings.

2.22.3 Communicative Outcome

Communicative outcome requires further study in the context of the Oral Interview (Clark 1982). This issue can only be resolved when testers specify communicative outcomes prior to the interview and when examinees are advised of and accept the validity of the desired outcomes. Pre-interview tapes, made by examinees and screened by testers, could guide the formulation of questions and situations prior to the interview.4
During the Oral Interview warm-up, as well, examinees and testers could negotiate communicative outcomes. Testers also could fine-tune interviews to examine and evaluate communicative competence as suggested by Sanders (1981).

2.23 Communicative Competence Caveat

Joseph Ricciardi (1981), in assessing the use of interviews, does not equate comprehension and use of linguistic structures with communicative competence. Yet the Oral Interview may still offer some insights to communicative competence in certain sociolinguistic situations. Ricciardi, however, advises caution with respect to the interview's predictive strength and mechanical application, specifically:

the entire range of problem areas that will be met in its application must be identified before the interview begins. In addition, the validity [of the Oral Interview] as an accurate judge of non-interview type oral language situations must also be ascertained before it is put into use.

It follows that descriptions of work environment and peer group speech acts, as well as validity studies, can guide interview developments.

2.24 Dutch Developments

The Dutch National Institute (DNI) for Educational Measurement constructed English, French and German oral proficiency tests based on daily living situations (Mets...
Rating difficulties, however, led the DNI to simplify its rating categories from six to three:

1. an intelligent and appropriate response with minor mistakes
2. an intelligible and appropriate response, but one which takes some effort on the part of the rater to understand
3. a response which is not forthcoming, not intelligible, or not appropriate in the context.

Next, the DNI decided to rate fluency and pronunciation four ways: (1) good, (2) adequate, (3) inadequate and (4) weak. Further developments by the DNI included a program which encompassed more clearly defined categories and ratings for intelligible responses lacking in vocabulary while, in addition, DNI raters strove to evaluate social and contextual appropriateness.

2.25 A Final Comparison

Compared with these recent evaluations and developments with respect to both oral proficiency and communicative competence, College Board Advanced Placement course objectives leave much to chance. For example, an Advanced Placement high school course would have the following objectives for students (CEEB 1984, 2):

A. Ability to understand spoken French in various unspecified conversational situations
B. French vocabulary sufficiently ample for reading newspaper and magazine articles, modern literature, and other non-technical writings without dependence on a dictionary
C. Ability to express oneself in French accurately and resourcefully in speech and writing with reasonable fluency.
These objectives, particularly those referring to spoken language, inadequately specify performance criteria and communicative needs, which are necessary to valid test construction and to communicative competence evaluation, respectively. With respect to oral proficiency and communicative competence, more concrete objectives which respond to real needs are required to evaluate foreign language speakers' abilities.

2.30 Communicative Competence

Test designers tend to specify desired linguistic manifestations before they completely specify communicative functions. Likewise they tend, prematurely, to consider pedagogical and logistical constraints. Communicative needs, however, exist independent from linguistic realizations and instructional programs. An instructional program is necessary only when a gap emerges between an individual's performance and that person's communicative needs. The basis of communicative performance testing development then, for Brendan J. Carroll (1980), is communicative needs analysis. Carroll mindfully calls his work an interim study to stress that development is continuing in the field of communicative performance testing.5

2.31 Needs Analysis

For Carroll (1980, 18-24), communicative needs analysis boils down to ten parameters and procedures. For
example, his needs analysis begins with participant identification, specifying language other than the target language and including the mother tongue. Carroll further specifies the individual's purpose: occupational, educational or social, or a combination. Communicative needs analysis also specifies the communicative setting, physical and psycho-social, while interaction is specified in terms of social relationships. Carroll's parameters and procedure have been implemented by a team developing this testing approach.

2.32.0 Test Development

Following his needs analysis, Carroll (1980, 18-71) presents, in detail, the procedure used to construct communicative test items. He bases his example on a profile of English for business studies.

2.32.1 Assessment Strategy

Carroll says tests cannot realistically measure speaking as an isolated language activity, oral proficiency alone that is. He questions the authenticity, validity elsewhere, of narrowly focused oral performance measurements. Nonetheless, he utilizes testing techniques which include the following:
1. information in the mother tongue (translation)
2. describing: experience/profession
3. transcoding: graphic information to speech
4. describing: an event just enacted (reporting)
5. presenting a mini-lecture: audience

Even though these techniques can focus on speaking proficiency and minimize interaction among speaking and the other three communicative modes: listening, reading, and writing, for Carroll at least, complete testing of communicative competence involves integrated language skills. His position, then, lends support to the indivisible competence hypothesis.  

2.32.2 Two-Tier Testing

Finally, Carroll proposes a two-tier communicative testing system. Tier One would be characterized by a basic test designed to evaluate realistic usage rather than real language use. Such a basic test would incorporate general skills and general contexts, not entirely suitable to all learners' communicative needs. Second tier tests would measure specific functions in particular content areas, but, because of the particulars, such detailed measurements would not form a basis for competence comparisons ranging across wide populations.

2.33.0 Closing the RACE

In order to meet demand in foreign languages, a scholastic as well as political—even universal impetus, communicative competence guidelines must be established.
Rather than proceed from pre-determined linguistic skills and standard testing techniques at our disposal, to build up to communicative competence and its evaluation, it is more desirable to begin with a theoretical if not descriptive framework for communicative performance, to set forth realistic communicative goals and testing guidelines and, then, on these bases, to construct instruments adaptable to the needs of language learners and evaluators. Such goals and guidelines, however flexible, will prove invaluable to communicative competence evaluation.

2.33.1 The RACE

Useful after test design, development and operation, in addition to before as goals, the RACE combines four serviceable criteria for evaluating communicative competence measurement instruments (Carroll 1980, 13-16):

1. Relevance of content based on analysis of needs, value in decision making
2. Acceptability—eliciting participant cooperation; respecting cultural contexts, religious and social concerns
3. Comparability—applicable over time, among individuals as well as among groups
4. Economy of time, effort and resources; capable of providing sufficient data for evaluation

It should be noted, here, that economical testing would utilize rapid subjective assessments based on carefully devised criteria, and scaled in comparison to objective test scores.
2.33.2 Revision

Carroll's interim study (1980) calls for consistent, continuous development and revision encompassing both tests and language instruction programs. The RACE encapsulates four desirable characteristics of communicative tests, a welcome change from standard theoretical and tactical positions. Whatever the approach, however, communicative competence evaluation requires constant validation efforts with respect to content and independent criteria.
CHAPTER 2

NOTES

1 See Section 1.13.21 for an Oral Interview validity study concluding in favor of partially divisible language competence; see Sections 2.13 and 2.32.1 for discussion supporting indivisible language competence.

2 Appended to Hendricks and others find interview topics divided into six groups (in Oller and Perkins 1980, 89-90): (1) present tense, (2) past tense, (3) future tense, (4) should or imperative, (5) conditional and (6) direct or indirect speech.

3 See Section 1.04, Note 2, for discussion of rating scale currency; see also Section 3.14.1, Note 2.

4 This would approximate Tier Two in Carroll's approach (1980); see Section 2.33.2.

5 Carroll's work (1980) exemplifies design, development, operation and assessment of communicative measurement instruments; then Carroll reviews communicative testing literature to date; appended to his work find a needs profile in English for business studies, test development data bases, sample rating scales and score correlation matrices for 156 Spanish speakers.

6 See Carroll (1984, 24), Figure 5, for streamlined specifications; see Carroll's APPENDIX I for a complete profile.

7 See Sections 1.13.1, 2.00 and 2.13 for further discussion of indivisible language competence.

8 Compare second tier tests envisioned by Carroll (1980, 7) with ILR job-specific interviews, Section 1.13.3, Note 7, and Section 1.14.2.
Numerous applications exist for oral proficiency and communicative competence test results. Entry, course and exit evaluations seem to frame language tests into suitable categories. First there are pre-course evaluations, where language tests are used for placement, for prediction and for measuring aptitude at or prior to the time students enter language programs. On-course evaluations, Carroll (1980, 73-84) suggests, may measure progress, facilitate diagnosis and provide feedback for formative course evaluation. At the end of courses, language tests may serve accreditation, prognosis and, again, course evaluation, this time summative, purposes.

Finnochiaro (1983), on the other hand, in a more traditional approach, distinguishes four kinds of language tests:

1. achievement or attainment tests
2. proficiency tests
3. diagnostic tests
4. aptitude tests

The first three, at least, characterize to some extent the measurement instruments described and evaluated in CHAPTER ONE of this work.
Here, however, I have divided applications for oral proficiency and communicative competence testing into two categories: academic and non-academic. Based on this division, I report how professional groups, schools and colleges have begun to apply oral and communicative measurement results, then how these results may influence decisions made in government and industry.

3.10 Academic Applications

This section includes studies and applications of oral proficiency measurements in the academic sphere. The presentation follows chronological developments in this area, which instructors should understand bear equally upon future evaluations to be made with respect to communicative competence.

3.11 University of Pennsylvania

Barbara Freed (1981) discussed studies and steps undertaken at the University of Pennsylvania to establish proficiency based language requirements. The faculty, there, administered College Board Language Achievement Tests as well as Oral Interviews. Measurement results figured in the faculty's development of performance standards which apply to foreign language learners at the post-secondary level.
3.12 Secondary School

Likewise, at the secondary level, according to Protase Woodford (1982), school teachers need a practical and valid means to measure foreign language students' speaking ability. At the same time, students and teachers, here, should strive to define proficiency goals that they can attain in high school foreign language courses. Teachers, both foreign language and bilingual, also need to develop their testing expertise in evaluating speaking proficiency and communicative competence.

3.13.0 University of South Carolina

Millstone (1983), in a paper presented to the Modern Language Association, announced several recent developments at the University of South Carolina.

3.13.1 Students' View

At first, according to Millstone, students at the University of South Carolina viewed the Oral Interview with anxiety when the French faculty established it as a degree requirement. Students felt that it represented a late-in-the-game barrier to their graduation. Simultaneously, however, aiming to assist students in achieving the required level of speaking proficiency, the University of South Carolina enhanced curricula in several languages.
3.13.2 Other Applications

The University of South Carolina also used proficiency tests in placing students in foreign language courses (Millstone 1983). Then, in addition to placement, the university used test results in predicting success in an international business internship program. This prediction, if successful, indicates limited validity of particular instruments for the purpose of selecting program candidates. It is still necessary, in applying test results to predicting success in either foreign language courses or in related programs, to consider the unpredictability of skill loss and even skill acquisition under different circumstances over time.

3.14.0 Project OPT

The purpose of this Oral Proficiency Testing project, commissioned by the U.S. Department of Education, was to investigate the feasibility of oral proficiency measurement at the college level, trained professors to administer Oral Interviews and initiated communication about oral testing among interested professors in a network of New England colleges and universities. Project OPT even generated oral proficiency tester certification requirements and, in the project report, distinguished interviewers from raters (Wing and Mayewski 1984, 17-21).1
3.14.1 Findings

The authors of this report concluded that there was no need for special conditions or equipment for individual interviews, beyond those existing and available in most foreign language departments. Tape recorders proved useful without their presence adversely affecting college level examinees. Wing and Mayewski did find, however, that the five to twenty minute interview had to be free of interruptions in order for it to produce a ratable speech sample.

Additional difficulties accompanied advantages in college level Oral Interview administrations. For example, professors encountered difficulty in suppressing their classroom teaching behavior during the interview. Further, concerning appropriate interviewing procedure, professors found interviews with their own students the most difficult. Whatmore, the ACTFL/ETS oral proficiency definitions were found to require revision to resolve possible ambiguities in the rating procedure.²

3.14.2 Summary

Regardless of the need for revision, Barbara H. Wing and Sandi F. Mayewski (1984, 10-12) summarized applications of the Oral Interview at the post-secondary level as follows:
1. placing students in language courses involving speaking skill
2. evaluating students' speaking proficiency within foreign language courses
3. selecting students for study abroad and monitoring the same for exit proficiency when they return from or complete their studies abroad
4. verifying proficiency required for:
   a. foreign language students in general
   b. foreign language majors in particular
   c. future foreign language teachers

3.15 Texas Teacher Certification

ACTFL helped establish a proficiency-based language instructor certification model for the State of Texas, where, according to Barbara Gonzalez (1984), a new state law in Texas will require prospective language teachers to pass an Oral Interview for certification in 1986 and afterward.

Gonzalez reported that most foreign language teachers surveyed (N: 142, in Texas, 1983) "wanted a description of the [oral proficiency] test and a list of the functions and notions covered." The foreign language teachers also expressed the need for funding to enhance their foreign experience, for new materials and for assistance reformulating objectives and learning new teaching methods. Further teachers' needs included informed access to more language topics and linguistic situations appropriate to communication in the target language and, finally, small classes assisted by competent personnel beyond the lone teacher. Although the new
certification requirements have not been applied to in-service teachers, teachers surveyed generally favored proficiency requirements, given adequate administrative, financial and logistical support in response to their needs.

3.16 Implications

Standardized proficiency requirements and testing could provide uniformity across departments, across campuses and across the nation, according to Wing and Mayewski (1984, 14). Lowe (1984, 39-40), in turn, lauds "the fact that testing drives curriculum . . . because teachers [who teach-to-the-test] teach performance strategies, not solely content." This would mark a shift in language instructors' focii from achievement and content to oral performance and communicative competence.

Proficiency testing, therefore, portends re-definition of the goals and specific language objectives of foreign language programs, with the addition of proficiency requirements, in particular for graduating college seniors and for prospective teachers. University course goals may in fact implement ACTFL Proficiency Guidelines while more courses and tests designed around speaking proficiency simultaneously may persuade students that communication in a foreign language is a realistic expectation (Wing and Mayewski 1984, 14-15). Oral proficiency measurement, by
whatever means, should play an important part in midterm and final examinations evaluating students in conversation courses taught at all levels (Lowe 1984, 39). Finally, instructors and students alike stand to benefit from instructional methods using the target language and promoting meaningful discourse among peers in communicative settings (Wing and Mayewski 1984, 15-16). 3

3.20 Non-Academic Implications

Industry, in addition to the United States government, has shown interest in oral proficiency and testing. The Exxon Foundation, for example, provided private funding for research, training and curriculum development at the University of Pennsylvania, the nation's first regional proficiency center (Heidinger 1983). Nonetheless, more private and government support is needed.

3.21 Applications

Qualified evaluators can appraise individuals' ability to communicate in a foreign language. Objective and standardized appraisals of this ability could contribute to employment decisions in industry, commerce, social services, travel and entertainment, as well as in government and other career fields (Wing and Mayewski 1984, 13).
3.22 Government Responsibility

Pardee Lowe, Jr. (1984, 39-40), for the United States government,

... underscores the nation's need for functional foreign languages, language ... which feeds, clothes, aids in travel, and beyond the initial stages, language capable of discussing differences between cultures, ... societies, [and] ... national policies.

Lowe urges adoption of the Oral Interview and its adaptation in the academic sector so that academic programs will conform to government goals, though this is not necessarily a major advantage to language instructors or to language learners, and so that the nation will have a common metric capable of measuring foreign language proficiency. By calling upon the academic community to train and test personnel to meet its needs, the government could reduce its responsibility for language training and testing.

3.30 Foreign Language Futures

Certainly oral proficiency testing, proficiency-based requirements and curricula call for further attention. Valid applications of communicative competence measurements, as well, require studies at the local level by schools, colleges and universities. Language instructors and adult language learners must keep abreast of current developments, beyond those I have examined here.
in order to define and evaluate their expectations and objectives, realistically, for both oral proficiency and communicative competence in foreign languages.
CHAPTER 3

NOTES

1See Sections 1.12.0 ff. and 1.14.5 for a discussion of the roles and training of testers.

2Appended to Wing and Mayewski (1984, 26-31) find, first, the ILR (government) Oral Proficiency Scale and, second, the ACTFL/ETS (academic) Oral Proficiency Scale; according to correspondence from Judith E. Liskin-Gasparro (17 May 1985), in a hand-written note to Paul A. Beaufait, the ACTFL/ETS scales are provisional and currently under revision; see Section 1.04, Note 1.

3ACTFL also publishes Applications of Oral Proficiency in Foreign Language Instruction (Hasting-on-Hudson, NY: ACTFL, 1985) for high school French, German, and Spanish teachers; in it, ACTFL reports adaptations made by teachers who participated in the Summer Proficiency Institute at Haverford College, Pennsylvania; this publication was mentioned in a memo (16 May 1985) inserted in a complimentary copy of Wing and Mayewski (1984) received by Dr. O. W. Rolfe, Department of Foreign Languages and Literatures, University of Montana, Missoula, MT 59812.
**APPENDIX A**

### Academic and Government Rating Scales

(from Liskin-Gasparro 1983)

#### ACADEMIC (ATTITUDES) RATING SCALE

<table>
<thead>
<tr>
<th>Scale</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>39.1.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** The table above represents a rating scale for academic attitudes, with ratings from 0 to 4 indicating different levels of agreement or disapproval with various statements. The specific items and their corresponding ratings are not fully visible in the image, but the table structure is clearly depicted.

---

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
<table>
<thead>
<tr>
<th>LEVEL 1</th>
<th>No Functional Ability in the Language</th>
</tr>
</thead>
</table>

In able to satisfy basic survival needs and listing few requirements. In order of immediate use of the very limited speech and some use of simple gestures, can ask directions, initiate and respond to simple statements, and maintain very limited face-to-face conversations. Within the scope of very simple interactive activities can understand simple questions and statements, including for example, directions or simple requests. When asked to do so, is able to participate after listening with limited concentration and short pauses. Most everyday activities require frequent and detailed guidance. Little development in stress and attention is evident.

<table>
<thead>
<tr>
<th>LEVEL 2</th>
<th>Able to satisfy basic survival needs and listing few requirements.</th>
</tr>
</thead>
</table>

...
<table>
<thead>
<tr>
<th>CRITERIA</th>
<th>FACTS</th>
<th>AGED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Criteria:**
- Relevant and specific for the intended audience.
- Consistent with the overall purpose of the document.
- Adequately supported by evidence or data.

**Facts:**
- Accurate and based on reliable sources.
- Easily verifiable by others.
- Complete and comprehensive.

**Aged:**
- Based on the most current information available.
- Relevant to the current context or situation.
- Up-to-date and not outdated.

The goal is to ensure that the information presented is accurate, relevant, and up-to-date, providing a clear and comprehensive understanding of the topic or issue at hand. This approach helps in making informed decisions and ensuring the effectiveness and relevance of the information provided.
<table>
<thead>
<tr>
<th>LEVEL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>3+</td>
<td>Able to speak the language with sufficient structural accuracy to perform basic functions in the use of language in social situations. Can speak in sentences with correct grammar and vocabulary, with pronunciation, vocabulary, and grammatical structures.</td>
</tr>
<tr>
<td>3</td>
<td>Able to use the language fluently and accurately on all levels, normally exhibiting all features of a native speaker's speech, including pronunciation, vocabulary, and grammatical structures.</td>
</tr>
<tr>
<td>2</td>
<td>Exhibits proficiency equivalent to that of a well-educated native speaker, with complete facility in the language and the ability to use it in all levels and all situations, including pronunciation, vocabulary, and grammatical structures.</td>
</tr>
<tr>
<td>1</td>
<td>Exhibits proficiency, equivalent to that of a well-educated native speaker, with complete facility in the language and the ability to use it in all levels, including pronunciation, vocabulary, and grammatical structures.</td>
</tr>
</tbody>
</table>

90

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
ILR Government Definitions: Oral Proficiency and ACTFL/ETS Provisional Speaking Descriptions

(ACTFL 1982)

ILR GOVERNMENT DEFINITIONS:

ORAL PROFICIENCY

NO PRACTICAL PROFICIENCY
S-0 Uses a few isolated words and phrases which have no practical application. Unable to participate even in a very simple conversation.

ELEMENTARY PROFICIENCY
S-1 Able to satisfy routine travel needs and minimum courtesy requirements. Can ask and answer questions on very familiar topics within the scope of very limited language experience used under normal circumstances. Using simple questions and statements, allowing for simple leven English interpretation 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 the language. Whole topics which are "very familiar" and elementary as to very considerably from individual to individual, anyone 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.

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 autobiography information. Can handle limited work requirements, needing help or 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 respond simply with some circumlocutions, assume, though often quite faulty, is intelligible, can usually handle elementary constructions quite accurately but does not have thorough or fluent control of the grammar.

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 so that the speaker 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.

DISTINGUISHED 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 one's personal and professional 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.

NATIVE OR BILINGUAL PROFICIENCY
S-5 Speaking proficiency equivalent to that of an educated native speaker. Has complete fluency in the language such that speech on all levels is easily accepted by educated native speakers in all of its features, including breadth of vocabulary and idiom, collocation, and pertinent cultural references.
THE ACTFL/ETS PROVISIONAL SPEAKING DESCRIPTIONS

The provisional descriptions for the lower end of the ILR scale were taught for the first time at the ACTFL/ETS workshop held under the sponsorship of a grant to ACTFL from the U.S. Department of Education entitled "Professional Development: Oral Proficiency Testing and Rating.

NOVICE—LOW
Unable to function in the spoken language. Oral production is limited to occasional isolated words. Essentially noncommunicative ability.

NOVICE—MID
Able to operate only in a very limited capacity within predefined areas of need. Vocabulary is limited so that necessary expressions can be understood in spoken sentences. Grammatical accuracy is limited to simple constructions, e.g., subject-verb agreement, noun-adjective agreement, some sound reflection. Vocabulary permits discussion of topics beyond basic survival needs, e.g., personal history, leisure-time activities. Is able to formulate some questions when asked to do so.

INTERMEDIATE—MID
Able to tackle some survival needs and some limited local demands. Some evidence of grammatical accuracy in basic constructions, e.g., subject-verb agreement, noun-adjective agreement, some sound reflection. Vocabulary permits discussion of topics beyond basic survival needs, e.g., personal history, leisure-time activities. Is able to formulate some questions when asked to do so.

INTERMEDIATE—HIGH
Able to handle basic survival needs and limited local demands. Developing flexibility in a range of circumstances beyond immediate survival needs. Shows some spontaneity in language production but fluency is very uneven. Can initiate and sustain a general conversation but has little understanding of the social constraints of conversation. Limited vocabulary range necessitates much hesitancy and circumlocution. The commoner tense forms occur but errors are frequent in formation and selection. Can use most question forms. While some word order is established, errors still occur in more complex patterns. Cannot sustain coherent structures in longer utterances or unfamiliar situations. Ability to describe and give precise information is limited. Averse to basic cohesive features, e.g., pronouns, verb inflections, but many are unreliable, especially in less immediate reference. Reaction time is longer in a series of short, discrete utterances. Articulation is comprehensible to native speakers used to dealing with foreigners, and can combine more phonemes with reasonable comprehensibility, but still has difficulty in producing certain sounds, in certain positions, or in certain combinations. Speech will usually be babbled. Still has to repeat utterances frequently to be understood by the general public. Able to produce some narration in either past or future.

ADVANCED
Able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations, including descriptions 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. Has a speaking vocabulary sufficient to respond to a variety of circumstances: abstract, technical, social, immigration, and wide areas of general conversation. Can generally satisfy most communicative needs and some demands for more detailed expression.

SUPERIOR
Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversation on practical, social, and occupational topics. Can discuss part-time interests and special fields of competence with reasonable ease. Vocabulary is broad enough that the speaker can handle most social situations; the general public usually understands the speaker with little difficulty. No more than occasional errors in pronunciation and sentence structure. Pronunciation is generally considered sufficiently accurate; does not have a foreign accent. Vocabulary is sufficiently large to handle most communicative situations, even the less immediate ones. Foreigners in general can understand the speaker with little difficulty. Pronunciation is generally considered sufficiently accurate; does not have a foreign accent. Vocabulary is sufficiently large to handle most communicative situations, even the less immediate ones. Foreigners in general can understand the speaker with little difficulty.
## APPENDIX C

### ILR Functional Trisection, ACTFL Provisional Guidelines: Generic Descriptions and German Descriptions

*(Lowe 1984, 42-49)*

<table>
<thead>
<tr>
<th>LEVEL</th>
<th>FUNCTIONS</th>
<th>CONTENT*</th>
<th>ACCURACY**</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLR Speaking level</td>
<td>Task accomplished. Attitudes expressed. Tone conveyed.</td>
<td>Topics, subject areas, activities and jobs addressed.</td>
<td>Acceptability, quality, and accuracy of message conveyed.</td>
</tr>
<tr>
<td>5</td>
<td>Functions equivalent to an Educated Native Speaker (ENS).</td>
<td>All subjects.</td>
<td>Performance equivalent to an ENS.</td>
</tr>
<tr>
<td>4</td>
<td>Can converse in formal and informal situations, create problem situations, deal with unfamiliar topics, provide explanations, describe in opinions, and hypothesize.</td>
<td>All topics normally pertinent to professional needs.</td>
<td>Nearly equivalent to an ENS.</td>
</tr>
<tr>
<td>3</td>
<td>Can converse in formal and informal situations, express facts, give instructions, describe, report on, provide narration about current, past, and future activities.</td>
<td>Practical, social, professional and abstract topics, particular interests, and special fields of competence.</td>
<td>Errors never interfere with understanding and rarely disturb the native speaker (NS).</td>
</tr>
<tr>
<td>2</td>
<td>Can create with the language: ask and answer questions, participate in short conversations.</td>
<td>Concrete topics such as own background, family, and interests, work, travel, and current events.</td>
<td>Understandable to an NS not used to dealing with foreigner.</td>
</tr>
<tr>
<td>1</td>
<td>No functional ability.</td>
<td>Everyday survival topics and courtesy requirements</td>
<td>Intelligible to an NS used to dealing with foreigners.</td>
</tr>
</tbody>
</table>

*May be no monosis.
**See also factor performance rating scales.

93

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Provisional Generic Description -- Speaking

Voice -- Low
Unable to function in the spoken language. Oral production is limited to occasional isolated words. Essentially, no communicative ability.

Voice -- High
Able to operate only in a very limited capacity within very predictable areas of need. Vocabulary limited to that necessary to express simple elementary needs and basic courteous formulae. Syntax is fragmented, inflections and word endings frequently omitted, confused or distorted and the majority of utterances consist of isolated words or short formulae. Utterances rarely consist of more than two or three words and are marked by frequent long pauses and repetition of an interlocutor's words. Pronunciation is frequently unintelligible and is strongly influenced by first language. Can be understood only with difficulty, even by persons such as teachers who are used to speaking with non-native speakers or in interactions where the context strongly supports the utterance.

Voice -- High
 Able to satisfy immediate needs using learned utterances. Can ask questions or make statements with reasonable accuracy only where this involves short memorized utterances or formulae. There is no real autonomy of expression, although there may be some emerging signs of spontaneity and flexibility. There is a slight increase in utterance length but frequent long pauses and repetition of interlocutor's words still occur. Most utterances are telegraphic and word endings are often omitted, confused or distorted. Vocabulary is limited to areas of immediate survival needs. Can differentiate most phonemes when produced in isolation, but when they are combined in words or groups of words, errors are frequent and, even with repetition, may severely inhibit communication even with persons used to dealing with such learners. Little development in stress and intonation is evident.

Intermediate -- Low
Able to satisfy basic survival needs and minimum courtesy requirements. In areas of immediate need or on very familiar topics, can ask and answer simple questions, initiate and respond to simple statements, and maintain very simple face-to-face conversations. When asked to do so, can formulate some questions with limited constructions and much inaccuracy. Almost every utterance contains fractured syn-
tax and other grammatical errors. Vocabulary inadequate to express anything but the most elementary needs. Strong interference from native language occurs in articulation, stress and intonation. Misunderstandings frequently arise from limited vocabulary and grammar and erroneous phonology. But, with repetition, can generally be understood by native speakers in regular contact with foreigners attempting to speak the language, Little decision in information conveyed owing to tentative state of grammatical development and little or no use of modifiers.

Intermediate -- Mid
Able to satisfy some survival needs and some limited social demands. Is able to formulate some questions when asked to do so. Vocabulary permits discussion of topics beyond basic survival needs such as personal history and leisure time activities. Some evidence of grammatical accuracy in basic constructions, for example: subject-verb agreement, noun-adjective agreement, some notion of inflection.

Intermediate -- High
Able to satisfy most survival needs and limited social demands. Shows some spontaneity in language production but fluency is very uneven. Can initiate and sustain a general conversation but has little understanding of the social conventions of conversation. Developing flexibility in a range of circumstances beyond immediate survival needs. Limited vocabulary range necessitates much hesitation and circumlocution. The commoner tense forms occur but errors are frequent in formation and selection. Can use most question forms. While some word order is established, errors still occur in more complex patterns. Cannot sustain coherent structures in longer utterances or unfamiliar situations. Ability to describe and give precise information is limited. Aware of basic cohesive features such as pronouns and verb inflections, but many are unreliable, especially if less immediate in reference. Extended discourse is largely a series of short, discrete utterances. Articulation is comprehensible to native speakers used to dealing with foreigners, and can combine most phonemes with reasonable comprehensibility, but still has difficulty in producing certain sounds, in certain positions, or in certain combinations, and speech will usually be labored. Still has to repeat utterances frequently to be understood by the general public. Able to produce some narration in either past or future.

Advanced
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. Has a speaking vocabulary sufficiently to respond simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite accurately but does not have thorough or consistent control of the grammar.

Advanced Plus
Able to satisfy most work requirements and show some ability to communicate on concrete topics relating to particular interests and special fields of competence. Generally strong in either grammar or vocabulary, but not in both. Weakness or unevenness in one of the foregoing or in pronunciation results in occasional miscommunication. Areas of weakness range from simple constructions such as plurals, articles, prepositions, and negatives to more complex structures such as tense, usage, passive constructions, word order, and relative clauses. Generally controls general vocabulary with some groping for everyday vocabulary still evident. Often shows remarkable fluency and ease of speech, but under tension or pressure language may break down.

Superior
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. Vocabulary is broad enough that speaker rarely has to grope for a word; accent may be obviously foreign; control of grammar good; errors virtually never interfere with understanding and rarely disturb the native speaker.

Provisional German Descriptions—Speaking

Novice — Low
Unable to function in spoken German. Oral production limited to occasional isolated words such as ja, nein, ich, Sie, Fritz (name), Frau, Lein. Essentially no communicative ability.

Novice — Mid
Able to operate only in a very limited capacity within very predictable areas of need. Vocabulary is limited to that necessary to express simple elementary needs and basic courtesy formulae such as Guten Tag/Morgen; Auf Wiedersehen; Sie ist . . . (name); war ist . . .; war ist das? Name, etc.Speakers at this level cannot create ori-
oral sentences or cope with the simplest situations. Pronunciation is frequently unintelligible and is strongly influenced by the first language. Can be understood only with difficulty, even by persons such as teachers who are used to dealing with nonnative speakers or in interactions where the context strongly supports the utterance.

Novice -- High

Able to satisfy immediate needs using learned utterances. There is no consistent ability to create original sentences or cope with simple survival situations. Can ask questions or make statements with reasonable accuracy only where this involves short memorized utterances or formulae. Vocabulary is limited to common areas such as colors, days of the week, months of the year, names of basic objects, numbers, and names of immediate family members — Vater, Mutter, Geschwister. Grammar shows only a few parts of speech. Verbs are generally in the present tense. Errors are frequent and, in spite of repetition, may severely inhibit communication even with persons used to dealing with such learners. Unable to make one's needs known and communicate essential information in a simple survival situation.

Intermediate -- Low

Able to satisfy basic survival needs and minimum courtesy requirements. In areas of immediate need or in very familiar topics, can ask and answer some simple statements. Can make one's needs known with great difficulty in a simple survival situation, such as ordering a meal, getting a hotel room, and asking for directions; vocabulary is adequate to talk simply about learning the target language and other academic studies. For example: Wie viel kostet das? Wo ist der Bahnhof? Ich möchte Ausflug. Ich habe eine Frage. Awareness of gender apparent (many mistakes). Word order is random. Verbs are generally in the present tense. Some correct use of predicate adjectives and personal pronouns (ich, wir). No clear distinction made between polite and familiar address forms (Sie, du). Awareness of case system sketchy; frequent errors in all structures. Misunderstandings frequently arise from limited vocabulary and grammar and erroneous phonology, but, with repetition, can generally be understood by native speakers in regular contact with foreigners attempting to speak German. Little precision in information conveyed owing to tentative state of grammatical development and little or no use of modifiers.

Intermediate -- Mid

Able to satisfy most routine travel and survival needs and some limited social demands. Can ask and answer questions on very familiar topics and in areas of immediate need. Can...
Initiate and respond to simple statements, and can maintain simple face-to-face conversation. Can ask and answer questions and carry on a conversation on topics beyond basic survival needs or involving the exchange of personal information, i.e., can talk simply about autobiographical information, leisure time activities, academic subjects. Can handle simple transactions at the post office, bank, drugstore, etc. Misunderstandings arise because of limited vocabulary, frequent grammatical errors, and poor pronunciation and intonation, although speakers at this level have greater vocabulary and/or greater grammatical and phonological control than speakers at Intermediate-Low. Speech is often characterized by long pauses. Some grammatical accuracy in some basic structures, i.e., subject-verb agreement, word order in simple statements (including adverbial and interrogative forms), present tense of irregular verbs and imperative of separable prefix verbs ('geben - gegeben'). Fluency is still strained but may be quite natural while within familiar territory. Is generally understood by persons used to dealing with foreigners.

Intermediate -- High

Able to satisfy most survival needs and limited social demands. Developing flexibility in language production although fluency is still uneven. Can initiate and sustain a general conversation on factual topics beyond basic survival needs. Can give autobiographical information and discuss leisure time activities. Most verbs are still in the present tense, more common past participles appear ('gegeben, gehehen, geschlafen'). Many mistakes in choice of auxiliary ('habe gegeben' with the present perfect). Past tense is attempted also with common imperfect forms ('haben, hatt, war'). Several high-frequency separable prefix verbs appear in the indicative (ich tanke mit). There is inconsistent coding of proper dative and accusative cases following prepositions in singular and plural. Attempts to expand discourse which is only accurate in short sentences. Frequently gropes for words. Comprehensible to native speakers used to dealing with foreigners, but still has to repeat utterances frequently to be understood by the general public.

Advanced

Able to satisfy routine social demands and limited school and work requirements. Can handle with confidence but not with facility most social and general conversations. Can narrate, describe and explain in past, present, and future time. Can communicate facts -- what, who, when, where, how much--and can explain a point of view, in an uncomplicated

Denotes an error characteristic of speakers at this level.
fashion, but cannot conjecture or coherently support an opinion. Can talk in a general way about topics of current public interest (e.g., current events, student rules and regulations), as well as personal interests (work, leisure time activities) and can give autobiographical information. Can make factual comparisons (e.g., life in a city vs. life in a rural area). Can handle work-related requirements, needing help in handling any complications or difficulties. Can make a point forcefully and communicate needs and thoughts in a situation with a complication (e.g., calling for help with a stalled car, losing traveler's checks). Has a speaking vocabulary sufficient to respond simply with some circumlocutions. Can be understood by native speakers not used to dealing with foreigners, in spite of some pronunciation difficulties. Good control of all verbs in present tense, past participles of most verbs, simple past tense of most irregular verbs, modal auxiliaries, most separable verbs and some reflexives. Double infinitives in main clauses may be attempted (mistakes are expected). Gender of high frequency words is mostly correct. Some inaccuracy in choice of prepositions as well as in distinctions between position and motion. Speaker is hesitant at times and gropes for words, uses paraphrases and fillers, uncomplexed dependent clauses, but mistakes are expected when sentences are joined in limited discourse.

**Advanced Plus**

Able to satisfy most school and work requirements and show some ability to communicate on concrete topics relating to particular interests and special fields of competence. Can narrate, describe, and explain in past, present, and future time. Can consistently communicate facts and explain points of view in an uncomplexed fashion. Shows some ability to support opinion, explain in detail, and hypothesize, although only sporadically. Can discuss topics of current and personal interest, and can handle most situations that arise in everyday life (see Advanced Level examples) but will have difficulty with unfamiliar situations (e.g., losing a contact lens in a sink drain and going to a neighbor to borrow a wrench). Nominally controls general vocabulary with some grogging still evident. Speaking performance is often uneven (e.g., strong in either grammar or vocabulary but not in both). Good control of most verbs in present and past tense and most imperative forms, irregular control of infinitive clauses with \textit{to} clausal sentences (with \textit{if} plus infinitive, \textit{that}, \textit{when}, \textit{where}, \textit{why}, and \textit{how} - and \textit{what} - compounds). Better control of prepositions and adjective endings but mistakes will occur. Control of dependent clauses. Distinguishes between subordinating and coordinating conjunctions and how

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Superior

Able to speak the language with sufficient structural accuracy and vocabulary to participate in most formal and informal conversations on practical, social, and professional topics. Can discuss particular interests and special fields of competence with reasonable ease. Can support opinions, hypothesize, and conjecture. May not be able to tailor language to fit various audiences or discuss highly abstract topics in depth. Vocabulary is broad enough that speaker rarely has to grope for a word; good use of circumlocution. Pronunciation may still be obviously foreign. Control of grammar is good. Sporadic errors but no patterns of error in tenses, cases, attributive adjectives, pronouns, most verbs plus preposition, dependent clauses, subjunctive II (present and past). Control less consistent in low frequency structures such as passive plus modals, the lassen construction, verbs plus specific prepositions (sprechen auf, sich halten an, sich über an), directional adverbs (hinweg, hinunter, herunter), double infinitives in dependent clauses (dass er das nicht hat machen sollen), varying degrees of competence in usage of idiomatic expression and slang. Errors never interfere with understanding and rarely disturb the native speaker.
APPENDIX D

Test of Spoken English Rating Scales: Overall
Comprehensibility, Pronunciation,
Grammar and Fluency
(TSE 1982, 11-13)

Overall Comprehensibility

0 - 90 Overall comprehensibility too low in even the simplest type of speech.

100 - 140 Generally not comprehensible due to frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary, and lack of grammatical control.

150 - 190 Generally comprehensible but with frequent errors in pronunciation, grammar, choice of vocabulary items, and with some pauses or rephrasing.

200 - 240 Generally comprehensible with some errors in pronunciation, grammar, choice of vocabulary items, or with pauses or occasional rephrasing.

250 - 300 Completely comprehensible in normal speech, with occasional grammatical or pronunciation errors in very colloquial phrases.

Pronunciation

0.0 - 0.4 Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be unintelligible.

0.5 - 1.4 Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be occasionally unintelligible.

1.5 - 2.4 Some consistent phonemic errors and foreign stress and intonation patterns, but speaker is intelligible.

2.5 - 3.0 Occasional nonnative pronunciation errors, but speaker is always intelligible.
Grammar

0.0 - 0.4 Virtually no grammatical or syntactical control except in simple stock phrases.

0.5 - 1.4 Some control of basic grammatical constructions but with major and/or repeated errors that interfere with intelligibility.

1.5 - 2.4 Generally good control in all constructions, with grammatical errors that do not interfere with overall intelligibility.

2.5 - 3.0 Sporadic minor grammatical errors that could be made inadvertently by native speakers.

Fluency

0.0 - 0.4 Speech is so halting and fragmentary or has such a nonnative flow that intelligibility is virtually impossible.

0.5 - 1.4 Numerous nonnative pauses and/or a nonnative flow that interferes with intelligibility.

1.5 - 2.4 Some nonnative pauses but with a more nearly native flow so that the pauses do not interfere with intelligibility.

2.5 - 3.0 Speech is as smooth and as effortless as that of a native speaker.
### Simplified Parameters and Procedures for Communicative Test Design

(Carroll 1980, 24)

#### Table: Parameters and procedures for test design

<table>
<thead>
<tr>
<th>Information Bank</th>
<th>Procedural Guide</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Participant identification</td>
<td>Broadly describe typical participant.</td>
</tr>
<tr>
<td>2. Purpose for language use</td>
<td>Describe main uses of English and classify under ESP headings: academic, occupational or social survival.</td>
</tr>
<tr>
<td>3. Events/activities</td>
<td>Choose the major events to be met with, and select several activities for each.</td>
</tr>
<tr>
<td>4. Instrumentality</td>
<td>Select media - listening, speaking, reading or writing, or multiple-mode combinations. Channels: face-to-face, tape, print, film, etc.</td>
</tr>
<tr>
<td>5. Socio-cultural</td>
<td>Specify social relationships, dialect and socio-cultural factors.</td>
</tr>
<tr>
<td>6. Performance levels</td>
<td>Using nine-point scale, give target levels of performance for each medium and multiple-mode.</td>
</tr>
<tr>
<td>7. Topic areas</td>
<td>Identify semantic areas for each specified event.</td>
</tr>
<tr>
<td>8. Language skills</td>
<td>Choose the skills necessary for carrying out the different activities at given target levels.</td>
</tr>
<tr>
<td>9. Language function/tone units</td>
<td>Indicate functions needed, and appropriate attitudinal tones, for those activities involving suitable person-to-person interactions.</td>
</tr>
<tr>
<td>10. Test format</td>
<td>Choose types of item for each activity - closed-ended, open-ended or restricted response, more RACE and authenticity.</td>
</tr>
</tbody>
</table>

* Caroll and others use the acronym ESP for English for Special Purposes.

** See Sections 2.33.0-1 for discussion of RACE.
SELECTED SOURCES


Ballard, Rod. 1985. Program Director, Test of Spoken English (TSE) and Speaking Proficiency English Assessment Kit (SPEAK) Programs, ETS: Princeton, NJ. Telephone, 26 April 1985.


ETT: ETT publications are arranged alphabetically by author or, in lieu of author, by title.


Mitchell, James V., ed. 1983. Tests in Print III. Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska.


Pacific Northwest Council on Foreign Languages: See PNCFL.


107


Woodford, Protase E. 1982. "Foreign Language and Bilingual Assessment." In Issues of Language Assessment. Edited by Stanley S. Seidner. [ED221023]