Development and Supervision of Teaching Assistants in Foreign Languages

Joel C. Walz
Editor

THOMSON
HEINLE

THOMSON
HEINLE

AAUSC
Development and Supervision of Teaching Assistants in Foreign Languages
Edited by Joel C. Waltz

Copyright © 2000 Heinle, a division of Thomson Learning, Inc.
Thomson Learning ™ is a trademark used herein under license.

Printed in the United States of America
3 4 5 6 7 8 9 10 06 05 04 03 02

For more information contact Heinle, 25 Thomson Place, Boston, MA 02210 USA.
or you can visit our Internet site at http://www.heinle.com

All rights reserved. No part of this work covered by the copyright hereon may be reproduced or used in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, Web distribution or information storage and retrieval systems—without the written permission of the publisher.

For permission to use material from this text or product contact us:
Tel 1-800-730-2214
Fax 1-800-730-2215
Web www.thomsonrights.com

ISBN: 0-8384-5124-1

BEST COPY AVAILABLE

Improving Inter-rater Reliability in Scoring Tests in Multisection Courses

Robert M. Terry
University of Richmond

At many colleges and universities lower-level foreign language courses are most frequently offered in multiple sections. Typical multisection courses have the following elements in common: course goals and objectives, syllabus, textbooks, tests (including final examinations), and a teaching corps often composed primarily of untrained graduate teaching assistants (TAs). Normally, the course goals and objectives, syllabus, and textbooks are determined either by a consensus of the teaching staff or by the coordinators of the course. Course tests and examinations may come from or be based on the test packet that accompanies the textbook, or they may be newly developed each semester or quarter by the instructional staff involved or by the course coordinators. A problem exists, however, with tests administered to all sections. Whether multisection courses are taught by TAs, part-time or adjunct faculty, or full-time faculty, and whether the tests from all sections are combined and graded by all instructors or individual instructors grade their own tests, the fundamental problem that arises is ensuring consistency among the instructional staff in scoring the tests.

This consistency in scoring is called *inter-rater reliability*, the "correlation between different raters' ratings of the same objects or performance"

(Henning, 1987, p. 193). While inter-rater reliability normally refers to scores on a given test that are independent assessments by two or more judges or raters, the term refers here to the correlation of scores on common tests among course sections, with each test being scored either by the TA of that section or randomly by one of the total group of TAs. High inter-rater reliability is important in order to ensure that course goals are being met and that student knowledge and performance are measured with a common yardstick.

Testing and Proficiency-Oriented Instruction

A major trend in current foreign language instruction is the shift in emphasis along the achievement/proficiency continuum. We have moved away from focusing on the knowledge of discrete grammar points and structures toward encouraging a more comprehensive, functional ability to carry out spontaneous, autonomous communicative tasks and exchanges in the target language in the four skill areas of reading, writing, listening, and speaking. Tests administered in those courses that claim to have such a communicative focus, however, regardless of their surface appearance, are often achievement tests that measure specific features of the language from a finite corpus of specific teaching materials. True proficiency tests, on the other hand, are performance-oriented and require the application of acquired knowledge to carry out communicative tasks. As Larson and Jones (1984, p. 116) note, "They are based on functional language ability and are not limited by a closed set of course materials nor constrained by instructional variables." Although many tests today include proficiency-oriented items that are contextualized and more comprehensive, that is, global, in scope, TAs may use them essentially for measuring student achievement. This is not contradictory, however, as Medley (1985, p. 35) so aptly points out:

> The fact that a curriculum is proficiency-oriented does *not* mean that there is no need for achievement testing. Quite the contrary, since achievement testing will be the primary means at the teacher's disposal for day-to-day assessment of student progress and instructional effectiveness. Tests will be designed to measure the specific features of the language the students are learning, and how well they are learning them. As a result, achievement will remain the principal determining factor in the measurement of progress and the assignment of grades.

Sections of a test containing items that are presented in a relatively traditional format, for example, fill-in-the-blank, multiple choice, sentence rewriting, cloze, and so on, contain essentially convergent items. Convergent items may be either discrete-point (focused) or in an integrative format scored by discrete points. This traditional format is most often used in evaluating the receptive skills (reading and listening), vocabulary, and basic knowledge of grammar. Those items or sections of the test that measure and evaluate the productive skills of speaking and writing are often more open-ended, global comprehension items that call for divergent answers (Omaggio, 1986, p. 315). Discrete-point scoring is objective, while the scoring of items with divergent answers tends to be subjective.

It is this latter category of subjective scoring that causes the most concern in ensuring high inter-rater reliability, since subjective scoring may be impressionistic, based on subjective reactions to responses to test items. Reactions can be positive toward the work of the student who shows strong communicative skills, who proves to be very creative, or who is very accurate. They can be negative toward the work of the student who, though he or she communicates well, possibly demonstrating a high degree of creativity, exhibits serious grammatical errors. As Magnan (1985, p. 130) cautions, "[W]e need to guard against judging all aspects of [a student's performance] in terms of only one dimension of it.... In grading, as in giving feedback, we should not let attention to errors in mechanics overshadow more communicative aspects."

The primary problem is, therefore, how to reduce subjectivity in evaluating student performance on divergent items without narrowing evaluation to a form of discrete-point scoring in which grammatical, lexical, and stylistic errors often overshadow other aspects of performance.

Errors of any sort and severity should not go unnoticed, since measuring and evaluating student performance should measure *all* aspects of the communicative act, whether communication takes place through writing or through speaking. Language is, after all, comprehensive; it is predicated on the appropriate combination and use of grammatical, lexical, stylistic, and sociolinguistic elements in order to convey intended meanings. A communicative act, even one marked by errors, is effective when, in fact, it communicates meaning. Nonetheless, we cannot let errors go unnoticed. Richards (1974, p. 49) wisely cautions: "If grammatically deviant speech still serves to communicate the speaker's intent, why should we pay further attention to it? Simply because speech is linked to attitudes and social

structure. Deviancy from grammatical and phonological norms of a speech community elicits evaluational reactions that may classify a person unfavorably." The same principle clearly exists for writing.

Scoring Techniques and Guidelines

Since the basic problem addressed in this chapter is how to ensure consistency among TAs in evaluating student performance on common tests, especially on divergent items, I will present many grading models for both writing and speaking. More important, however, than the model selected is the manner in which it is used in the multisection course.

The following imperative is offered: establish clear guidelines for assessing and scoring such tests. Supervisors and TAs should draw up these guidelines *before* administering tests. The guidelines should contain clearly written descriptions that specify the variety of types of performance that TAs can expect according to the level of the course, the material that has been covered, the students' level of linguistic sophistication and knowledge, and what the students are expected to know at this particular stage of language learning. It is only logical to assume that the descriptions will be based on experience with students at a given level of study. Such a set of guidelines will establish a priori the common yardstick against which all students will be measured.

TAs must be trained in using the guidelines since, in reality, no one student's test will match letter for letter the descriptions of any one level of performance. TAs must realize that each test will exhibit certain key traits that are indicative of performance for each level. For this reason, the guideline descriptions must be specific enough to discriminate between levels of performance, identifying the key traits that are manifested at each level, without being so specific that every mistake that could be made is enumerated. They should also be general enough to include all anticipated varieties of typical student performance.

The operative word in the preceding paragraph is *trained*. The problem of ensuring inter-rater reliability is exacerbated when the instructional personnel in core multisection courses are made up of novice TAs "whose appointment rests primarily upon the survival needs of an understaffed system of higher education" (Murphy, 1991, p. 130). While TAs must fulfill the requirements to earn their graduate degree, they must also be trained in foreign language pedagogy, which adds yet another course to their already

overcrowded program of study. Such methods courses are crucial for instilling a high degree of professional competence in tomorrow's group of foreign language teachers. As Murphy (1991, p. 141) points out, "The new TA suffers from a flawed educational system in which he/she enters graduate school deprived of requisite background knowledge. The problem is two-dimensional: (1) undergraduate programs in the liberal arts are often inadequate for personal development needs and (2) professional or specialist training in the rudiments of teaching is missing."

The AATF (Murphy & Goepper, 1989), AATSP ("AATSP," 1990), and ACTFL ("ACTFL," 1988) have identified general competencies for foreign language teachers, among which are found:

- The teacher who possesses the Basic level of competence will "know how to prepare *instruments with which to diagnose and evaluate the skill areas* of speaking, listening, reading, and writing as well as a knowledge of culture" (Murphy & Goepper, 1989, p. 21; my italics).
- Indicators of program consistency with the goal of instruction include: "coursework and experience in devising *appropriate testing techniques*" ("ACTFL," 1988, p. 77; my italics).

The AATF's "Teaching of French: A Syllabus of Competence" (Murphy & Goepper, 1989) recommends both a preservice and an advanced methods course. For TAs especially, the preservice course should offer these prospective teachers "some theoretical concerns and a variety of techniques with which to enter the profession and develop a routine" (Murphy & Goepper, 1989, p. 21). The variety of techniques must include practice in the evaluation and grading of student performance, since evaluation techniques must be clearly understood before the first test is actually administered and graded. Through such intensive training the common yardstick can be introduced, practiced, and understood, thereby ensuring consistency in scoring throughout multisection courses.

Solution 1: Holistic Scoring

The essential purpose of all testing is to determine levels of student performance (knowledge) based on the comparison of performance against the standards established for a course. One type of guideline that can be drawn up for assessing student performance is a global scale in which comprehensive, descriptive criteria are written for each of the expected levels

of performance. A holistic (global) scale is based on an overall, total *impression* of student work, since certain components in free, creative responses cannot be quantified as discrete-point items because there can be no clear-cut anticipated response. In evaluating student performance on highly creative test items and activities, there is no substitute for the judgment of the evaluator in determining the overall impression of the response, that is, the balance that students have achieved between grammatical accuracy and meaningful communication. In spite of this necessity, such judgment is still subjective. Holistic criteria help reduce the subjective nature of scoring by providing clear descriptions of levels of performance against which student performance can be determined. The descriptive criteria should be written for each different divergent test item as well as for each test since student abilities are expected to increase and improve over time. Even if the same test is used in subsequent years, the criteria should undergo periodic revision and refinement, resulting in a more precise reflection of the various levels of performance expected in the program, and the development of scales that are clearer and easier to use by the TAs, "enabling them to assign a grade with as little arbitrariness as is humanly possible" (Johnson, 1983, p. 17).

Holistic scoring is used in scoring the Advanced Placement (AP) Examinations of the College Board. We must realize, however, that AP tests are administered to high school students who are at an advanced level. The expectations of performance and the resultant scoring scales advocated by the College Board far surpass the pragmatic reality of beginning- and intermediate-level college and university courses in which students are only starting to reach the level of typical AP students. Nonetheless, these scoring scales are worth examining, since they can be adapted for use at a level commensurate with the expectations of performance for lower-level students.

Standards for each question on the AP examination have an associated grading scale that is designed to allow readers to make distinctions among answers. "The scales — usually from 0 to 9 or 0 to 15 — avoid the problem, on the one hand, of too few points which allow only coarse distinctions and, on the other hand, of too many points which require overly refined, often meaningless, discriminations. The grading standards guarantee that no matter when a candidate's answer is read or by whom, it will, in all probability, receive the same grade" (Johnson, 1983, p. 3). *This* is inter-rater reliability.

The standardization of scoring is carried out by having a supervisory group of readers score sample answers individually, then compare their scores. After discussing the sample answers, they reach a consensus on the grade that the sample answers should receive. Significant examples are chosen for each grade level, that is, level of performance, defining the exact standards by which all examinations are to be graded, providing examples of the application of these standards, and ensuring that there is a common understanding of how the standards are to be applied (Johnson, 1983, p. 17).

The sample answers selected by the supervisory group are then scored, analyzed, and discussed with the group of scorers. Much attention is paid to those borderline cases that fall just to one side or the other of the critical line dividing an acceptable performance from one that does not suggest achievement acceptable at the AP level. Then, a group of preselected essays is distributed and individually scored. The process of constantly comparing scores and discussing reasons for assigning a given score leads to a clear understanding of the gradations of the scale by the scorers. Once this understanding is achieved, formal scoring can begin. This same type of training program is used for scoring student-taped oral performance on the speaking section of the examination.

As I noted above, similar training programs should exist for TAs who teach multisection courses, whether they are experienced teachers or neophytes. In the case of TAs, it is crucial that training in evaluation go hand in hand with teacher training. Many TAs have no prior teaching experience and rely solely on impressions or recollections of how they themselves were taught and tested. Optimally the evaluation training period for TAs should take place before instruction in the course actually begins. The training program should familiarize the TAs with the goals and objectives of each course and give them practice in scoring tests, thereby ensuring a clear understanding of the application of the scoring system.

Writing

The scoring standards for the essay-writing section of the AP French examination are found in Table 1. "The score that is given to any particular exam is determined largely by the student's use of language as measured against the scale given" (Johnson, 1983, p. 18).

Table 1

Advanced Placement Scoring Standards: Writing

Demonstrates Superiority	9	*Strong* control of the language: proficiency and variety in grammatical usage with few significant errors; broad command of vocabulary and of idiomatic French.
Demonstrates Competence	8 7	*Good* general control of grammatical structures despite some errors and/or some awkwardness of style. Good use of idioms and vocabulary. Reads smoothly overall.
Suggests Competence	6 5	*Fair* ability to express ideas in French: correct use of simple grammatical structures or use of more complex structures without numerous serious errors. Some apt vocabulary and idioms. Occasional signs of fluency and sense of style.
Suggests Incompetence	4 3	*Weak* use of language with little control of grammatical structures. Limited vocabulary. Frequent use of anglicisms which force interpretations on the part of the reader. Occasional redeeming features.
Demonstrates Incompetence	2 1	*Clearly unacceptable* from most points of view. Almost total lack of vocabulary resources, little or no sense of idiom and/or style. Essentially gallicized English or *charabia.*
Floating Point		A one-point bonus should be awarded for a coherent and well organized essay or for a particularly inventive one.

From Johnson, 1983, p. 18.

Reschke (1990, p. 101) points out that there are two aspects to global assessment of *any* linguistic performance, regardless of modality, that are important in assessing writing (as well as speaking) skills: (1) the shift in the focus of the evaluation from the usual preoccupation with student errors to

what the student does *well* and *correctly;* and (2) the degree to which the student *succeeds* in expressing and communicating his or her ideas. Reschke has modified the College Board's nine-point AP scale slightly to make it applicable for use by foreign language teachers. He provides two different scales: a basic, intuitive scale that serves both as an initial and as a final check in the evaluation process (see Table 2), and a complementary articulated scale (see Table 3) with more complete descriptions of each of the five proficiency levels it identifies.

Table 2

Reschke's Holistic Essay Grading Scale: Basic Scale — Range and Minimal Description

Upper Half	
9	Demonstrates superior writing skills
8	_____
7	Demonstrates strong writing skills
6	_____
5	Demonstrates competent writing skills
4	_____
Lower Half	
3	Suggests incompetent writing skills
2	_____
1	Demonstrates incompetent writing skills

From Reschke, 1990, p. 102.

Table 3

Reschke's Holistic Essay Grading Scale: Articulated Scale

Upper-Half Responses

9 to 8 Demonstrates superiority through outstanding control of the language with regard to syntax, grammar, idiomatic usage, and vocabulary. The student makes few significant errors and demonstrates a broad command of the language and obvious fluency. The difference between an 8 and a 9 is one of degree.

7 to 6 Demonstrates strong command of the target language with, however, some grammatical inaccuracies and errors and some awkwardness of expression. Shows good, although not always accurate, use of vocabulary and idioms. Errors do *not* detract from the generally clear demonstration of competence. The difference between a 6 and a 7 is one of degree (quality, fluency).

5 to 4 Demonstrates good to acceptable use of the language and suggests that the candidate is basically competent. The student makes occasional serious grammatical and syntactic errors and has a less impressive range of vocabulary and idioms than a student in the category above. There are occasional signs of fluency in the written work. Recurring doubt about the competence of a student lowers the score to a 4.

Lower-Half Responses

3 to 2 Weak use of the language suggests incompetence. The composition displays numerous errors and frequently uses anglicisms and/or English syntax and thought patterns. The composition contains sentences that paraphrase or essentially repeat what has been stated earlier, lists activities and places or things in series without giving reasons, and/or forces interpretation on the part of the reader. The lack of an occasional redeeming feature, such as the correct use of advanced grammatical constructions and vocabulary, tends to lower the score to a 2. (Getting a simple sentence grammatically correct now and then is *not* a sufficiently redeeming feature.)

1 Clear demonstration of incompetence. The student has little or no sense of syntax and has very few vocabulary resources. The content of the student's written work is essentially incomprehensible Germanized English.

Additional Comments:

a. One point is *subtracted* if the essay or composition does not address the assigned topic.

b. One point is *subtracted* if the essay or composition is poorly organized *or* is substantially shorter than called for (i.e., less than 90% of the assigned length).

c. One point is *added* if the essay or composition is especially well organized *and* well written.

d. No more than two points are deducted from any essay or composition.

e. In case of doubt about what score to assign to an essay or composition (a high 6/low 7 or a strong 7/weak 8), the spelling is carefully looked at. If it is obviously phonetic and poor (many errors), the lower score is assigned.

From Reschke, 1990, p. 103.

For our purposes, both the rationale and the principles that are the bases for creating such scales are what is important, not the specific wording as illustrated in the tables. It is clear that students in lower-level classes are not asked to write compositions or essays that involve high-level stylistic features. Often the TA cannot assess the value of the content given the autobiographical nature of many topics. Nonetheless, the student is demonstrating a developing writing skill along with a knowledge of the rudiments and formal aspects of the target language, and the impact of the entire writing sample must be considered in evaluation.

The principles of such scales as those illustrated in Tables 1, 2, and 3 can be readily adapted to suit the content and linguistic levels of the students, even students in the second semester of a beginning-level language class. It should be obvious that the wider the scale (the more ratings there are within each level), the more subjective the scoring becomes. With a range of 1–9, it is relatively easy to decide between a 5 or a 4, or even across levels between

a 7 and an 8. However, if the scores were to range from 1 to 20, with five possible scores in each level, it would be extremely subjective to decide whether a composition should receive a 16 or a 17. The narrower the scale, the more effect the floating bonus or penalty points would have, since a bonus/penalty of 1 point could move a composition from a 6 to a 7 on the AP scale (7 to 8 on the Reschke scale), thereby recognizing and rewarding those elements that contribute to the overall positive impression of the composition. Conversely, penalty points could lower a 4 to a 3 (Reschke scale) and thus affect the score of the student who wrote a grammatically accurate composition that exhibited poor organization, did not address the topic, was shorter than called for, or demonstrated other problems.

For beginning-level students, the holistic scale can be reduced even more, if the evaluation is to indicate a general impression and not detailed scoring of writing performance. Such a limited scale is found in the Virginia Standards of Learning, Cumulative Assessments, French II, Writing (Virginia Department of Education Standards of Learning: French, Spanish, German, and Latin, 1988, p. 4). (See Table 4.)

It is extremely important to react to *what* the student has said, not only to *how* it was said. Such reactions prove that the message is as important as the means of expressing it, in other words, that appropriate communication has taken place between a writer and his or her audience. Simply because student performance has been evaluated on a test and a grade has been assigned is no reason to think that the test is an end in itself. When samples of student writing have been evaluated, it is not unrealistic to ask students to revise their work. The subsequent revisions can be counted as homework, quizzes, or extra credit. Through encouraging rewriting and revising, TAs underline the necessity for clear communication and indicate that the writing *process* is as important as the written *product.*

The choice of the particular scale to be used is best determined by weighing several factors:

1) The level of the students
2) The amount of training in writing they have received
3) Expectations of performance
4) The weighting of the written section of the test with respect to the remainder of the test
5) The degree of refinement needed in order to assess student performance accurately.

Table 4

Limited Holistic Scale (General Impressions)
Virginia Standards of Learning: Cumulative Assessment,
French II, Writing

4	Can communicate a message in declarative, negative, and interrogative sentences. Errors in vocabulary, syntax, and mechanics are not consistent and do not interfere with intelligibility. They are able to recombine vocabulary and structures from the prompt. Most verbs may appear in the present tense. Past and future time may be expressed. Where sample is a letter, the appropriate date, salutation, greeting, and closing are included.
3	Can communicate most of the message intelligibly, but some errors in grammar, syntax, and mechanics interfere with the meaning. Where sample is a letter, date, salutation, greeting, and closing are included.
2	Can communicate some of the message, with minimum intelligibility. The message is greatly confused due to frequent and consistent errors in syntax, vocabulary, and mechanics.
1	Can communicate virtually none of the message intelligibly.

Virginia Department of Education, Cumulative Assessment, French II, Writing, 1988, p. 4.

We can expect more from students as their abilities increase with continued study of the language. With training in the development of writing skills and the elements and principles that make up "good writing," we can expect students to demonstrate a broader range of knowledge. Students, in turn, should expect the evaluation of their written work to be more detailed, following the more refined (maybe even more demanding) scoring guidelines that will be commensurate with their level of training and abilities.

Speaking

It is significantly more difficult to evaluate oral than written performance, since speaking is much more transitory than writing, unless it is captured on tape. With writing, the evaluator can read and reread. With speaking, the evaluator can hear the message only once. Taping can help to solve this problem, but introduces others: wearing earphones and listening to many different students over and over again is fatiguing, and the tapes themselves are often of poor technical quality. All of this makes evaluation extremely difficult and tedious.

Nonetheless, tape-recorded speaking test sections are administered on the AP examination. In one section, students hear questions or directions that establish a situation and must then respond to each situation with an appropriate answer. Each response is scored using a 4-point scale (see Table 5).

Even in scoring this relatively short section of the test, however, there are hazards: sometimes it is difficult to determine whether the student has really understood the question; sometimes the difference between a "major" and a "minor" error is unclear.

In the other section of the AP speaking test, students see a sequence of pictures illustrating a story that they are then asked to tell or interpret within a given time. This lengthier section is scored using a scale similar to the 9-point scale used for scoring the essay portion of the examination (see Table 6).

Table 5

Holistic Scoring for Questions and Directions
Advanced Placement — Speaking

4 points:	1) A correct answer to the question, delivered with excellent to good pronunciation, correct grammar, and considerable fluency.
	2) A longer, more elaborate answer to the question, but with a minor error or two in grammar, pronunciation, or usage.
3 points:	1) A correct answer to the question with fair pronunciation and intonation, perhaps a minor grammatical error or two, and some awkwardness in usage or delivery.
	2) A longer, more elaborate answer, with not more than *one* major grammatical error.
2 points:	A correct answer to the question, with less than fair pronunciation and intonation, delivered haltingly and/or with one or two major flaws in grammar or usage.
1 point:	1) An answer given in very faulty French, with little control of either grammar or pronunciation. The student is unable to express his thought with any competence.
	2) A comprehensible answer that shows that the students did not entirely understand the question.
	3) A response in which a major part of the answer is missing or not complete (two-part question, for example).
0 points:	1) An answer indicating total failure to understand the question.
	2) An answer so fragmented as to be incomprehensible.
	3) An answer such as "Je ne sais pas," "Je ne comprends pas," or any similar effort to evade the problem posed.
	4) No answer.

Johnson, 1983, p. 24.

Table 6

Holistic Scoring: 9-point Scale
Advanced Placement — Speaking

Demonstrates Superiority	9	*Strong* control of the language: excellent grammatical and idiomatic usage; broad command of vocabulary, and obvious ease of expression. No significant grammar or pronunciation errors.
Demonstrates Competence	8 7	*Good* control of the language, with some grammatical accuracies or some awkwardness of expression. Good intonation and use of idiom and vocabulary. Few glaring errors of grammar or pronunciation.
Suggests Competence	6 5	*Fair* use of language without numerous serious grammatical errors but with a less impressive range of vocabulary and idiom and less good pronunciation and intonation. Occasional signs of fluency.
Suggests Incompetence	4 3	*Weak* use of language with serious errors. Restricted vocabulary and knowledge of idioms and/or frequent use of anglicisms or sentences which force interpretations on the part of the reader. Some redeeming features.
Demonstrates Incompetence	2 1	*Unacceptable:* few vocabulary resources, little or no sense of idiom or French style, glaring weakness in pronunciation and grammar.

Johnson, 1983, p. 24.

The College Board recognizes the difficulty of using their holistic scoring scale and recommends an analytic scoring scale to double-check the grade given. (Analytic scoring is discussed later in this chapter.)

Because of the nature of the AP examination, such scoring criteria may prove to be too detailed for use with lower-level students. A very simple 4-point scale (see Table 7), similar to one used for writing, has been created for speaking in the Virginia Standards of Learning, Cumulative Assessment,

French II, Speaking (Virginia Department of Education Standards of Learning: French, Spanish, German, and Latin, 1988, p. 4).

Table 7

Simplified 4-Point Scale
Virginia Standards of Learning: Cumulative Assessment,
French II, Speaking

4	Can communicate a message intelligibly. Answers are in complete sentences, including simple interrogative and negative structures. Common and regular verbs are used in the present tense with some degree of accuracy. Some errors may occur, but these do not interfere with the message.
3	Can communicate most of the message intelligibly, but errors may cause some misunderstanding. Most answers are in complete sentences, including simple interrogative and negative structures. Common and regular verbs are used in the present tense, but some are misconjugated. Vocabulary limitations, grammatical errors, and weak pronunciation may cause some difficulty in communication, but do not interfere with the basic message.
2	Can communicate some of the message with minimal intelligibility, but errors cause frequent misunderstandings. Simple declarative, negative, and interrogative sentences are attempted, but most structures have fractured syntax. Verbs are used in the present tense, but most forms are misconjugated. Problems in vocabulary, grammar, or pronunciation sometimes interfere seriously with the basic message.
1	Communicates very little of the message intelligibly. Every sentence is marked by long pauses and serious errors, garbled syntax, or lapses into English.

Virginia Department of Education, Cumulative Assessment, French II, Speaking, 1988, p. 4.

It should be obvious that holistic scoring of speaking ability, even with scoring scales and tape recording, is still impressionistic. Nothing can be formally marked for correction or feedback. Errors and comments can only be noted as the mistakes are made, which interferes with listening, or after the fact, when specific errors are more difficult to remember.

Scoring scales do, however, provide less subjective guidelines for evaluating oral performance. Here, as with writing, the vital role of evaluation training sessions can be seen. Concentrated practice in evaluating sample tapes will help ensure that scorers understand the scoring system and its application. Thus, TAs will have guidance in determining what to listen for when rating the students' level of performance and in determining their grade.

Although the Oral Proficiency Interview (OPI) also calls for holistic scoring, it was never intended to be used in an academic setting as a means of evaluating student performance in speaking, since its very purpose is to determine a *proficiency* level, not *achievement.* There are several logistical problems encountered in administering the OPI in a classroom setting:

1) It should not be administered by someone who knows the person being interviewed.
2) It should be administered in a one-on-one situation.
3) It should last from 15 to 25 minutes.

Proficiency ratings *must not* be used to determine achievement or a grade. These aspects of formal proficiency interviews preclude their use in a typical classroom testing situation.

Solution 2: Analytic Scoring

According to Perkins (1983), cited in Omaggio (1986, p. 265), analytic scoring "involves the separation of the various features of a composition into components for scoring purposes." Analytic scoring offers more objectivity than holistic scoring in assessing student performance because it is more focused: the categories of language use to be evaluated are spelled out, and descriptions of performance levels within each category are provided.

In order to ensure standardization and consistency in scoring, each of the categories of language use must be clearly defined. It can be assumed that TAs understand terms such as "grammar," "vocabulary," "content," "fluency," "organization," "mechanics," "pronunciation," and so forth. Yet, to

ensure that the focus of each of the categories is clearly understood, each term should be defined, since the categories represent the components of student performance that are to be evaluated. Each category, furthermore, should be weighted in reference to the degree of importance that it carries in the test items being evaluated.

Writing

Table 8 illustrates a complete analytic scoring scale for writing in a beginning-level French class. (I wrote this analytic scale myself for evaluating a test item that involved answering a letter from an imaginary pen pal. Since the letter was seeded to elicit particular grammatical forms and structures that had been studied in the course, the grammar category is more specific than the other categories.)

As Table 8 suggests, one of the most attractive features of such a scoring scale is the grid that is used in indicating performance in each of the five categories. Students should receive a copy of the analytic scoring criteria in advance. They will then know what is expected of them and how to interpret the evaluation of their written work. They can readily see where their strengths and weaknesses lie and can, over time, visualize their progress with subsequent evaluated samples of their writing.

It is a simple matter to convert the total of the scores in the various categories by converting them to a scale of 100. If a 100-point scale is routinely used for grading in class, TAs simply multiply the total earned in the five categories of the writing sample by 4. The score can also be weighted as to its relative importance on the entire test if the writing sample is only one section of a longer test.

It must be pointed out again that the specific categories, definitions, and descriptions of levels of performance should be spelled out before grading student papers. Furthermore, the same analytic scale should not be used throughout the course, since both the categories and descriptions should change with the widening scope of course content. There is nothing sacred about having five categories as in Table 8; the number and types of categories should be determined by the particular emphasis put on the development of writing skills in the course. Similarly, the descriptions should become more refined as course content increases.

Table 8

Analytic Scoring

	5	4	3	2	1
Grammar					
Vocabulary					
Mechanics					
Fluency					
Relevance					

COMMENTS: _____

_____

GRAMMAR: Use of grammatical elements, i.e., various parts of speech: correct pronouns (subject, object, reflexive, stressed), verb persons and tenses, adjective agreement, appropriate use of articles, correct genders, appropriate negative elements, etc.

5 Excellent use of grammatical elements; very limited errors in gender and adjective agreement; correct pronoun substitutions (both subject and object); correct use of articles (definite, indefinite, partitive); widely varied use and correctness of verb tenses; errors are relatively insignificant and do not hinder comprehension.

4 Very good use of grammatical elements; few errors in gender and adjective agreement; verb tenses are limited, primarily in the present, but some effort at using other tenses; errors in verb forms but not serious enough to hinder comprehension; correct but limited use of pronoun substitutes.

3 Satisfactory use of grammatical elements; some significant errors but overall impression of text is affected by errors; noticeable errors in pronoun usage, genders, verb forms and tenses, adjective agreement and position, and article usage.

2 Unsatisfactory use of grammatical elements; too many serious errors hinder comprehension; significant amount of anglicized French in constructions; extremely limited pronoun substitutes

with incorrect forms or position; verb tenses limited to present with little or no effort at using other tenses; genders are often incorrect; adjectives show limited or no agreement; many errors in article usage.

1 Totally unsatisfactory use of grammatical elements; the severity of the errors obstructs comprehension of most of the text; strictly limited to present tense, and even then with serious errors; no grasp of genders and adjective agreement; articles are used haphazardly.

VOCABULARY: Appropriate lexical items, variety of types of lexical items.

5 Exceptional range of vocabulary; subtleties and idiomatic expressions are used appropriately, giving a sense of strong control of lexicon. Vocabulary elements go far beyond routine elements suggested by the task/stimulus.

4 Good range of vocabulary; awareness of subtleties is demonstrated although with some errors; some extraordinary vocabulary elements included not expected to be found in task/stimulus and used appropriately.

3 Limited range of vocabulary; predominantly copies vocabulary from stimulus or uses very routine vocabulary, at times inappropriately; significant errors in choice of certain items.

2 Extremely limited vocabulary; even items expected to be found in task or provided in stimulus are inappropriate with incorrect spelling or use.

1 Shows no grasp of appropriate vocabulary; serious errors in word choice; serious misspellings.

MECHANICS: Appropriate use of pronoun substitutes, varied sentence structures (including simple, compound, and complex structures), logically sequenced writing, appropriate use of cohesive elements (adverbs of time, conjunctions, pronouns).

5 Excellent control of a variety of structures: a variety of sentence types; excellent use of cohesive elements, including appropriate pronoun substitutions for both subjects and objects; writing is appropriately sequenced, illustrated with the use of time elements and other connective elements.

4 Strong control of structures: good sentence variety, not limited to simple, affirmative, active, declarative, sentences (SAAD); appropriate but limited use of pronoun substitutes; limited use of cohesive elements and time words.

3 Adequate control of structures; most sentences are SAADs with one or two attempts at compound/complex structures; limited use of pronoun substitutes, some inappropriate or incorrect; very limited use of sequencing elements.

2 Poor control of structures; text is limited to SAADs with much repetition of words rather than appropriate pronoun substitutes; even short simple sentences are often incorrect.

1 Demonstrates virtually no control of mechanics, including appropriately structured simple sentences; no evidence of cohesion and coherence in text.

FLUENCY: The amount of information provided, i.e., does the student go beyond what is called for in the task/stimulus and contribute further information or comments? How inventive and/or creative is the writing sample? What is the degree of risk-taking seen in the writing sample?

5 Student goes far beyond the task/stimulus and contributes additional information; creativity is evident and is in general appropriate. Student clearly demonstrates risk-taking in going beyond what was called for; text reads very smoothly.

4 Above average work; student goes beyond the task/stimulus but is hesitant in taking many risks; some additional comments and reactions expressed, most of which are acceptable and logical; evidence of creativity in responding and some inventiveness; text reads smoothly but with some awkwardness; writing sample is complete, i.e., adequately responds to the task/stimulus.

3 Limited fluency demonstrated; student basically responds to stimulus or performs on task without going beyond giving what is called for; some evidence of creativity, but errors impede comprehension when student attempts to go beyond the task; relatively limited amount of content, but what is there is appropriate; text is generally awkward and jerky.

2 Very limited evidence of creativity or inventiveness; student simply copies cues from the stimulus, often inappropriately; writing sample

is extremely short and incomplete; text is very awkward, jerky, and disconnected.

1 No evidence of creativity; text is totally uncreative; many stimulus questions and comments are left unanswered; stimulus text is simply copied and poorly at that; writing sample is entirely too brief and incomplete; text is virtually unreadable.

RELEVANCE: Student responses and reactions are relevant to the stimulus questions or comments.

5 Student responses and comments are totally appropriate and relevant; additional comments are consistent with the context of the task/stimulus; all of the writing sample is on task.

4 Student responses are in general appropriate and relevant but some extraneous, irrelevant information is given; most of writing sample is on task.

3 Student responses are often inappropriate due to misreading or lack of comprehension of task/stimulus.

2 Most student responses are irrelevant and incomplete, based on serious misinterpretation of task/stimulus cues.

1 Student shows no grasp of relevance due to strong lack of comprehension of task/stimulus text. Responses and comments are essentially "off the wall" with little meaning (such responses and comments are not creative in nature, but due to faulty comprehension).

Speaking

Analytic scoring again provides a less subjective manner of evaluating student oral performance. The categories to be evaluated are spelled out with a rating scale for the levels of performance for each category. TAs can use several different types of analytic scales, each of which can be adapted to fit a specific learning, teaching, or testing situation.

Bruschke (1989, p. 18) provides three different scales, two of which are based on the proficiency functional trisection of accuracy, content/context, and function (see Table 9 and Table 10). In her third scale (see Table 11) we find more specific categories and a more detailed description of each. What is particularly interesting in the scoring chart in Table 11 is the relative importance of each of the seven categories and the weight assigned to each: vocabulary, functions, and accuracy receive the highest weight of 5, fluency and pronunciation receive a weight of 3, and reaction/appropriateness and creativity/recombination receive a weight of 2.

Table 9

Evaluation of Oral Proficiency (I)

		1 2 3 4 5	
Function	Can't use language to communicate needs and ideas; has little functional ability; gropes for every word	_ _ _ _ _	Uses language to communicate needs and ideas; has good functional ability at his/her level of proficiency; language flows
Content/ Context	Has very limited vocabulary; uses vocabulary inappropriate to topic(s)	_ _ _ _ _	Has good command of vocabulary appropriate to topic(s)
Accuracy	Has poor word and sentence structure; has incomprehensible pronunciation	_ _ _ _ _	Has good word and sentence structure; has good pronunciation

Bruschke, 1989.

Table 10

Evaluation of Oral Proficiency (II)

		1	2	3	4	5
Function	Uses language to communicate needs and ideas; has functional ability at his/her level of proficiency; speaks at a normal pace	—	—	—	—	—
Content/ Context	Uses vocabulary appropriate to topic(s)	—	—	—	—	—
Accuracy	Uses correct word and sentence structure; pronunciation does not interfere with communication	—	—	—	—	—

Bruschke, 1989.

Table 11

Evaluation of Oral Proficiency (III) — Bruschke

I. *Vocabulary within Context*

0 1 2 3 4 5

minimal extensive

II. *Functions/Use of Language*
(i.e., give information, enumerate/describe, ask questions, express likes/dislikes)

0 1 2 3 4 5

few many

III. *Accuracy in Use of Basic Structures*
(i.e., subject–verb, noun–adjective agreement, basic word order, negation)

0	1	2	3	4	5
very poor, interferes with communication			few errors, does not interfere with communication		

IV. *Fluency*

0	1	2	3
none	groping, slow	occasionally fluent	confident, language flows

V. *Pronunciation/Intonation*

0	1	2	3
no effort to use target lang. sounds	poor, greatly interferes with communication	occasionally interferes	does not interfere with communication

VI. *Reaction/Appropriateness of Response*

0	1	2
no reaction	sometimes appropriate	appropriate

VII. *Creativity/Recombination of Learned Material*

0	1	2
no attempt	some attempt	frequent attempts

COMMENTS: _____

_____

_____

_____

Bruschke, 1989.

Pino (1989, p. 492) also presents two analytic scales in which the assignment of a score for the speaking test is greatly simplified (see Table 12 and Table 13). She also defines the five categories used in the evaluation. In Table 12, she presents a college or high school version that is appropriate for lower-level students, and in Table 13, a scale for more advanced students.

Table 12

Oral Language Rating Scale I

Categories	A+	A	B	C	D	F	Notes
Communication	40	37	34	31	28	25	
Accuracy	20	18	16	14	12	10	
Fluency	10	9	8	7	6	5	
Vocabulary	20	18	16	14	12	10	
Pronunciation	10	9	8	7	6	5	

Communication: Did you understand what was said to you? Are you talking about the right thing? Can you be understood despite errors? Have you conveyed your idea?

Accuracy: reasonable to inadequate grammatical correctness

Fluency: flow vs. hesitation

Vocabulary: adequate vs. inadequate

Pronunciation: good to bad

Pino, 1989, p. 492.

Table 13

Oral Language Rating Scale II

Categories	A+	A	B	C	D	F	Notes
Communication	25	23	21	19	17	15	
Accuracy	25	23	21	19	17	15	
Fluency	15	13	11	9	7	5	
Vocabulary	20	18	16	14	12	10	
Pronunciation	15	13	11	9	7	5	

Communication: purpose clearly conveyed for an average performance to purpose creatively and sensitively conveyed for an outstanding performance

Accuracy: grammatical correctness, especially _____ [teacher supplies grammatical features]

Fluency: amount of speech

Vocabulary: adequate for the purpose for an average score to advanced/new vocabulary for an outstanding score

Pronunciation: merely comprehensible to nativelike

Pino, 1989, p. 492.

It is again interesting to note the relative weight that Pino has given to categories on the two scales, especially the fact that *communication* is weighted the heaviest on Oral Scale I, while it is weighted equal to *accuracy* on Oral Scale II. Indeed, the effectiveness of the communicative effort is of utmost importance in oral communication. In speaking, the other participants can ask for repetition or clarification if there is misunderstanding or lack of comprehension, or if verification is needed; whereas with writing, the intended audience, a reader or readers, is not present and therefore cannot guide the writer by indicating problems with comprehension.

For nonnative teachers of another language, one of the most helpful criteria to bear in mind in evaluating speaking is the statement found in the ACTFL Proficiency Guidelines (1986, p. 2): "[T]he speaker can generally be

understood even by interlocutors *not accustomed to* dealing with speakers at this level" [my italics]. In other words, one should ask, "Can I put myself in the place of a native speaker and understand what I have just heard?" With experience, we foreign language teachers tend to "understand" much more student language than the native speaker would understand, since we are accustomed to dealing with speakers at this level. In evaluating speaking performance, however, we cannot falsify student ability by rating it too high simply because we understand what the student is attempting to say. For inexperienced TAs especially, it is essential that they learn to hear their students objectively. Obviously, the same criterion can be applied to writing: "Writing is understandable to natives not used to the writing of nonnatives" (ACTFL, 1986, p. 5).

Table 14 offers another analytic scale for oral grading procedures, one developed by Hirsch and Thompson (1989, p. 24). This scale provides two different formulae for evaluating oral performance. Each formula gives different weighting to the categories. Note that Formula One does not evaluate content. Whatever categories and descriptions may be used, the relative differences in weight are based on the realistic expectations for student performance at a given level of study and can easily be varied.

Table 14

Oral Grading Procedures

GRAMMAR

A: Usage of required grammar concepts is almost perfect in given context.

B: Makes some grammar mistakes which generally would not affect meaning (i.e., agreements, partitive vs. definite article, wrong past participles, etc.).

C: Makes more serious mistakes which could give unintended meaning (i.e., conjugation, tense inconsistency, word order mistakes).

D: Meaning frequently obscured by grammar mistakes.

E: *"Épouvantable."*

VOCABULARY

A: Conversant with vocabulary required by given context.
B: Makes some vocabulary mistakes which generally would not affect meaning (i.e., wrong gender, wrong preposition [*finir à* + inf.]).
C: Makes more serious mistakes which could give unintended meaning (i.e., wrong gender, incorrect word choice, mangled words).
D: Meaning frequently obscured because of inadequate mastery of vocabulary.
E: *"Épouvantable."*

FLUENCY

A: No more than a normal, "thoughtful" delay in formulation of thoughts into speech.
B: Hesitates longer than necessary to find the right word.
C: Narrative somewhat disjointed because of pauses.
D: Painful pauses make speech hard to follow.
E: *"Effrayant."*

PRONUNCIATION

A: Demonstrates a knowledge of correct pronunciation and intonation; makes very few mistakes.
B: Some mispronunciation, but meaning is still clear.
C: Pronounced foreign accent which requires extra-sympathetic listening.
D: Meaning frequently obscured because of poor pronunciation.
E: *"Épouvantable."*

CONTENT

A: Displays communicative ease within a given context.
B: Says more than the strict minimum.
C: Situation handled adequately though minimally.
D: Says less than adequate minimum.
E: Situation handled only partially or in totally unsatisfactory manner.

Formula One	*Formula Two*
Grammar: ___ × 7 = ___	Grammar: ___ × 5 = ___
Vocabulary: ___ × 6 = ___	Vocabulary: ___ × 5 = ___
Pronunciation: ___ × 4 = ___	Pronunciation: ___ × 3 = ___
Fluency: ___ × 3 = ___	Fluency: ___ × 2 = ___
	Content: ___ × 5 = ___

A	4.5 – 5.0
B	4.0 – 4.4
C	3.5 – 3.9
D	3.0 – 3.4
E	2.5 – 2.9

Hirsch & Thompson, 1989, p. 24.

Conclusion

To improve consistency when TAs evaluate their students' writing and speaking in ways that tend to call for subjective grading, this chapter has provided several guidelines and samples of scoring techniques. Whether holistic or analytic scoring and any of the scales presented are chosen, five points must be remembered:

1) All TAs who will score common tests should go through an intensive training period that will familiarize them with the techniques for scoring student work, allow for discussion of these techniques, provide sample texts for evaluation, and examine why selected samples were given certain scores, thereby ensuring consistency in the application of the scoring procedure. Such direction is vital for ensuring consistency among TAs, that is, for ensuring inter-rater reliability.
2) In multisection courses, it is extremely important to have clearly defined scoring criteria for all tests to ensure equal expectations and equivalent results on common tests that are scored by a number of different TAs.

3) Scoring scales should be created for each test administered. A common scoring scale for tests that cover an entire academic year is highly inappropriate.
4) Scoring criteria should include *all* types of items on the test, whether they are scored by discrete points or by holistic/analytic methods. These guidelines should be created *before* the test is administered.
5) For holistic scoring, descriptions for each level of performance should be general enough to include all varieties of anticipated student performance yet specific enough to give guidance in discriminating both quality and quantity of work between the ranges of levels of performance typical of the student population in the entire course. Categories and descriptions for levels of performance should be realistic, attainable, and commensurate with the students' level of study and degree of linguistic sophistication.

Consistency in grading provides the equitable evaluation of student performance across sections in multisection courses, especially in those courses taught by relatively inexperienced TAs. Such consistency can be improved — if not ensured — through rigorous training sessions in which TAs examine a variety of scoring scales and techniques, and practice scoring numerous samples of student performance in both speaking and writing. The scoring techniques suggested here will help reduce the gap between the totally impressionistic evaluation of student writing and speaking and the relatively objective evaluation of the various elements of effective language use.

Works Cited

"AATSP Program Guidelines for the Education and Training of Teachers of Spanish and Portuguese." *Hispania* 73 (1990): 785–94.

ACTFL Proficiency Guidelines. Yonkers, NY: ACTFL Materials Center, 1986.

"ACTFL Provisional Program Guidelines for Foreign Language Teacher Education." *Foreign Language Annals* 21 (1988): 71–82.

Bruschke, Dorothea. Personal communication and handout for "Teaching and Testing for Oral Proficiency." Wisconsin Education Association Council Meeting, Madison, WI, 1989.

Henning, Grant. *A Guide to Language Testing: Development, Evaluation, Research.* Cambridge, MA: Newbury House Publishers, 1987.

Hirsch, Bette G. & Chantal P. Thompson. *Ensuite: Cours intermédiaire de français.* Instructor's Manual. New York: Random House, 1989.

Johnson, Leonard W. *Grading the Advanced Placement Examination in French Language.* Princeton, NJ: College Entrance Examination Board, 1983.

Larson, Jerry W. & Randall L. Jones. "Proficiency Testing for the Other Language Modalities." *Teaching for Proficiency: The Organizing Principle.* Ed. Theodore V. Higgs. ACTFL Foreign Language Education Series. Lincolnwood, IL: National Textbook Company, 1984: 113–38.

Magnan, Sally Sieloff. "Teaching and Testing Proficiency in Writing: Skills to Transcend the Second-Language Classroom." In Omaggio, 1985: 109–36.

Medley, Frank W., Jr. "Designing the Proficiency-Based Curriculum." In Omaggio, 1985: 13–40.

Murphy, Joseph A. "The Graduate Teaching Assistant in an Age of Standards." *Challenges in the 1990s for College Foreign Language Programs.* Ed. Sally Sieloff Magnan. AAUSC Issues in Language Program Direction 1990. Boston: Heinle & Heinle Publishers, 1991: 129–49.

_____ & Jane Black Goepper (Ed.). *The Teaching of French: A Syllabus of Competence.* The Report of the Commission on Professional Standards, the American Association of Teachers of French. *AATF National Bulletin* 14 (October 1989). (Special issue.)

Omaggio, Alice C. *Teaching Language in Context: Proficiency-Oriented Instruction.* Boston: Heinle & Heinle Publishers, 1986.

_____ (Ed.). *Proficiency, Curriculum, Articulation: The Ties That Bind.* Middlebury, VT: Northeast Conference on the Teaching of Foreign Languages, 1985.

Perkins, Kyle. "On the Use of Composition Scoring Techniques, Objective Measures, and Objective Tests to Evaluate ESL Writing Ability." *TESOL Quarterly* 17 (1983): 651–71.

Pino, Barbara Gonzalez. "Prochievement Testing of Speaking." *Foreign Language Annals* 22 (1989): 487–96.

Reschke, Claus. "Global Assessment of Writing Proficiency." *Realizing the Potential of Foreign Language Instruction.* Ed. Gerard L. Ervin. Report of Central States Conference on the Teaching of Foreign Languages. Lincolnwood, IL: National Textbook Company, 1990: 100–11.

Richards, Jack C. "Error Analysis and Second-Language Strategies." *New Frontiers in Second Language Learning.* Ed. John Schumann & Nancy Stenson. Rowley, MA: Newbury House Publishers, 1974.: 32–53.

Virginia Department of Education Standards of Learning: French, Spanish, German, and Latin. Richmond, VA: Virginia Department of Education, 1988.