THE FRENCH REVIEW, Vol. LII, No. 4, March 1979 *Printed in U.S.A.*

Testing Communicative Ability*

by Albert Valdman and Marvin Moody

FOLLOWING SANDRA SAVIGNON'S PIONEERING STUDY,[1] foreign language (FL) educators have, increasingly, placed greater weight on the learner's attainment of some communicative ability, even at beginning levels of instruction. In this article, we discuss the testing of communicative ability at a relatively low level and describe an instrument that attempts to measure that skill within the context of a large, multi-section university beginning French program. We believe, however, that most features of the test of minimal communicative ability described here are adaptable to other teaching situations.

1. Current approaches and existing tests of speaking ability

One may, of course, question the need for formal instruments that test communicative ability. Might one not, instead, evaluate that skill on the basis of daily classroom performance? That solution has at least two major drawbacks. First, students are not likely to take seriously any teaching objective that is not tested in a formal and rigorous manner. Second, we have discovered that instructors' in-class evaluation of communicative ability is relatively unreliable. That judgment tends to be highly influenced by students' performance in other areas. Current research[2] indicates that communicative ability is determined by strategies of language acquisition distinct from many conscious mechanisms. It is a set of capacities independent from manipulative skills related to the development of narrow linguistic competence and must be measured by an independent instrument.

Available procedures for the testing of speaking proficiency are unsuitable

* We would like to express our gratitude to the many associate instructors in the Department of French and Italian at Indiana University who, between 1972 and 1977, have administered the test described in this paper. We wish to acknowledge particularly the special contribution of Betsy Kerr Barnes, currently Assistant Professor of French at Bowling Green University, and Beverly Tice. They jointly evaluated the reliability of scoring of IUFCAT and provided a general assessment of its reliability.

[1] Sandra J. Savignon, *Communicative Competence: An Experiment in Foreign-Language Teaching* (Philadelphia: Center for Curriculum Development, 1972).

[2] For example, Stephen Krashen, "The Monitor Model for Adult Second Language Performance," in M. Burt, H. Dulay, and M. Finocchiaro, eds., *Personal Viewpoints on Aspects of ESL* (New York: Regent Publishing Company, 1978) assigns language learning to conscious, monitored processes and language acquisition to a different set of psychological operations.

for use in a large program (ours enrolls more than 1500 students each year and is staffed exclusively by relatively inexperienced graduate student associate instructors). The oral interview developed for use at the Foreign Service Institute (FSI) of the Department of State is designed to measure a higher level of proficiency than that attained by beginning college students, and requires highly trained personnel to administer and score it reliably. Similarly, the MLA Cooperative and the Pimsleur tests[3] measure proficiency attained at the end of one year of college FL instruction. These instruments also suffer from serious shortcomings from a theoretical standpoint. First, they give preponderant emphasis to the evaluation of pronunciation accuracy, a skill that contributes relatively little to communicative ability. Second, on the basis of these tests it is difficult to separate a subject's communicative ability from such contaminating factors as general knowledge, inventiveness, intelligence, memory, and imagination. Third, both tests are fairly lengthy and their scoring in a reliable manner requires highly trained personnel. Finally, a more serious drawback is apparent from the very title of these two instruments: French *speaking proficiency* test. Communicative ability is an interactive, two-way skill that requires participants to decode as well as to encode messages. Yet both the MLA and the Pimsleur tests judge productive ability independently of receptive skills.

Valid evaluation of communicative ability requires a procedure that simulates use of language in an authentic situational context. The judgment must bear on the success of the communicative act and its semantic and pragmatic content rather than external well-formedness and accuracy. Fundamental to the preparation of any instrument for the testing of communicative ability is the clear recognition of the lack of direct relationship between analytical knowledge of discrete linguistic elements (words, grammatical rules, phonemes, etc.), or the ability to manipulate these discrete elements in artificial exercises, and the ability to make use of these elements in engaging in meaningful communication interactions. In this regard, the FSI directed interview constitutes a valid test of communicative ability. However, as was pointed out, it was designed to measure that skill at a very high level (the "passing" level of linguistic proficiency in the U.S. Foreign Service, S-3, represents competence attained usually after four years of college-level study) and is unsuitable for administration at the end of one semester of study.

In her trial of a pilot beginning French course stressing communicative ability, Savignon devised a suitable test providing evaluation of that skill in a variety of modes, but her procedure was designed for use only within the context of the experiment, and it does not pretend to serve as a model for the testing of communicative ability in a standardized manner in large programs such as ours.

[3] *MLA Cooperative Foreign Language Tests* (Princeton, N.J.: Educational Testing Service, 1964); Paul Pimsleur, "Testing Foreign Language Learning," in Albert Valdman, ed., *Trends in Language Teaching* (New York: McGraw-Hill, 1966), pp. 175–214.

2. The Indiana University French Communicative Ability Test (IUFCAT)

Faced with the lack of a suitable instrument for the evaluation of minimal communicative ability, yet recognizing the urgency to assess that skill, we set out in 1972 (before the availability of Savignon's report) to develop a test which would provide valid and reliable evaluation and which could be administered effectively in a large multi-section FL program. The test needed to meet the following requirements:

1. Valid measurement of communicative ability, as opposed to pronunciation accuracy or spoken proficiency;
2. Simulation of genuine conversational interaction in the target language;
3. Ease of administration and general economy;
4. Reliability of scoring on the part of relatively inexperienced teaching personnel.

The IUFCAT in its current form consists of four sections: part I, pictorially cued responses; part II, personal questions; part III, question formulation; part IV, situational responses.

The pictorially cued-response section of the test consists of six questions which are presented on tape. Students have ten seconds to respond to each question, the answer to which is provided by a drawing contained in the test form distributed to students. For example, students may hear *A quelle heure est-ce que Jacques s'est couché hier soir?* and are to respond using the information provided by the drawing of a clock which shows, say, 11:00. In part II, personal questions, students again hear six questions, but no cue is provided. A typical question from this section would be *Où habitent tes parents?* In the third section of the test, students are asked to formulate questions based on information provided in the form of three declarative sentences. For example, they may hear (a) *Il fait beau.* (b) *Il pleut.* (c) *Il neige.* The expected response is of course *Quel temps fait-il?* or any acceptable variant.

Each response is worth 8 points and is evaluated along three parameters: semantic and pragmatic appropriateness (0–3), grammaticality and correct form of lexical items (0–3), and fluency and accuracy (0–2). The first of these parameters refers to semantic content and reflects whether the response answers the question adequately. For example, an error in tense would result in the loss of 1 point, as would a mistake in the choice of subject pronoun. Grammaticality is judged independently of appropriateness and deals solely with the correctness of the morphology or syntax of the response. For example, a student might provide the following question formulation in response to the same item given above: *Quel temps faisait-il?* While this response is not altogether appropriate, it would be rated "3" on grammaticality. Pronunciation errors which result in grammatical or lexical errors are accounted for in this category. For example, if a student pronounced the final *s* of *ans* in *Il a dix-huit ans*, this would be considered a lexical error. The third parameter, fluency, is a global evaluation.

The evaluation scheme we have adopted raises the issue of the relevance of judgments of well-formedness and pronunciation accuracy in the assessment of communicative ability. Sandra Savignon's testing procedure involves primarily the parameters of communication of information (comprehensibility) and fluency of expression. On the other hand, the FSI scoring system requires judgment of relative grammaticality and, to a lesser degree, of pronunciation accuracy. In the face of a growing trend toward reducing of objectives which favors reading comprehension, it is important to assert the legitimacy of the goal of language instruction for communicative ability in general (i.e., non-special purpose). But communicative ability can be only a secondary objective and, realistically, only a minimal level can be attained. Given the special circumstances of classroom language instruction, we can only aim at the *simulation* of genuine communicative acts. It is therefore quite proper that the special – if not artificial – type of communicative ability gained in the classroom assign high value to language form and expression. There is also ample evidence that in the acquisition of a FL in a natural setting, as in immigrant speech, grammatically deviant though pragmatically adequate utterances are highly stigmatized by native speakers. While sensitivity to grammatical deviance varies from community to community, native speakers do seem to expect a high level of correctness from learners who have acquired the language by formal training. Our evaluation scheme, which allots half of the total points to semantic and pragmatic appropriateness and to fluency, and half to grammatical well-formedness and to pronunciation accuracy, appears fully compatible with the nature and objectives of the general FL course.

To give scorers a firmer base for judgments, we have attempted a description of the various points in the three evaluation scales:

Appropriateness

3 Totally appropriate response
2 Use of inappropriate person, number, or tense; or response that is partially inappropriate in a minor way
1 Response that is partially inappropriate in some major way, but that still demonstrates at least partial comprehension of the situation
0 Totally inappropriate response, or no response

Well-Formedness

3 Perfect or nearly perfect response, containing no more than one lexical error, i.e. mispronounced word
2 Somewhat deviant response, containing one grammatical error (morphology or syntax) or two lexical errors
1 Somewhat deviant response, not interpretable by native speaker
0 Very deviant error, not readily interpretable by native speaker; or no response

Pronunciation Accuracy and Fluency

2 No hesitation; absence of errors at the phonemic level and of gross, typical phonetic inaccuracies, e.g. gliding of final [e] and [o]
1 No hesitation; or response containing no more than one phonemic error or

gross phonetic inaccuracy
0 Hesitant response; or response containing two or more phonemic errors or gross phonetic inaccuracies

We illustrate the application of our evaluation scheme to various actual student responses to part II (personal questions). In response to *Qu'est-ce que vous allez faire samedi soir?* typical answers and scores were:

	Appropriate	Well Formed	Fluency Accuracy
(a) Samedi soir, je suis allé [alleʸ] au cinéma.	2	3	1
(b) Je vais . . . au . . . cinéma [sineʸma].	3	3	0
(c) Je suis étudié samedi soir [swɑʳ].	2	2	1

Response (a) contains the wrong tense and a gross phonetic error. Response (b) is appropriate, since in context the present may be used to refer to near future events, but too hesitant. Response (c) contains the past instead of the future and a morphological error; *soir* is produced with a retroflex [r] and accompanying backing of the vowel, a typical pronunciation error on the part of American learners.

Responses to the question *Combien est-ce qu'il faut payer pour voir un film ici?* posed other types of scoring problems:

	Appropriate	Well Formed	Fluency Accuracy
(d) On va payer cinquante cents [sã] pour voir un [œ̃n] film.	2	3	1
(e) On doit payer cinq [sẽk] francs.	3	2	2
(f) Il faut payer [peji] *two* dollars.	3	2	1
(g) Je n'ai pas vu [vu] les [leʸ] films.	1	3	1
(h) Il faut payer cinq francs.	3	3	2

Answer (d) contains a wrong modal, and the pronunciation of *un* as [œ̃n] is interpreted as a typical American phonemic error rather than a lexical error. In (e) the mispronunciation [sẽk] is scored as a lexical error, since it does not reflect typical American errors at the phonological level. In (f) the use of the English word *two* constitutes a lexical error. Response (g) poses problems of interpretation; the scoring assumes that the student meant *Je ne vais pas d'ordinaire au cinéma et je n'ai pas l'occasion de voir des films, donc je ne sais pas combien il faut payer*. While it is fluent, accurate, and well formed, it poses severe problems of comprehension in this situation and is judged as highly deviant.

Part IV, which we are currently developing, requires students to produce one or more sentences appropriate to a situation described in English. For example, for the situation: "You are in a café and you would like to order drinks for yourself and a friend" students might provide one of the following possible responses:

(a) Garçon, apportez un café pour mademoiselle et une bière pour moi.
(b) Mademoiselle, je prends une limonade et lui une tasse de thé.
(c) Je voudrais du coca-cola et ce monsieur un jus orange.

In addition to pragmatic and semantic appropriateness, grammatical well-formedness, and pronunciation accuracy and fluency, responses are judged on the basis of amount of information communicated. For each situation, one point is awarded for inclusion of one of a pre-determined number of informational elements. The situation given as illustration would require conveying a minimum of four informational elements: (1) term of address to waiter or waitress; (2) drink requested by speaker; (3) identification of other person; (4) drink requested for other person. Thus, response (a) would receive a score of "4," as would response (b), since the pronoun *lui* is sufficient in context; response (c), which lacks the term of address, would be scored only "3."

3. Characteristics of the IUFCAT

3.1 *Practicality*

The IUFCAT is relatively easy to administer to large groups of students if language lab facilities are available. Since the test lasts only 12 minutes, the time required to evaluate each tape is not unreasonable, requiring on the average 15–18 minutes (including the time required to put the tape on the machine, fill in information on the grading sheet, etc.). A related advantage is the relative ease of scoring. As noted earlier, most of our instructors are relatively inexperienced and it is thus necessary to provide them with training in the evaluation of this type of test. Such training is carried out in two ways. During a three-day orientation period that precedes the beginning of the academic year and the week before the final speaking test, the course director meets with the instructors as a group. Grading procedures are outlined and illustrated by playing a tape of sample student responses which are evaluated by the course director. The rationale for each judgment is explained. Then an actual test tape from the previous semester is played which instructors grade on sample grading forms. Evaluations are then discussed. In addition, during the actual scoring of the end-of-semester exams, the course director and other supervisory personnel make a random spot check of graded tapes. If it is found that a given evaluator's judgments vary significantly from those of the checker, the latter works with the former on an individual basis until the difficulty has been resolved. In order to make the evaluations more objective, instructors do not grade the tapes of their own students.

3.2 *Validity*

A test is considered valid if it measures what it claims to measure. One can judge the face validity, or content validity, by inspecting the test to

satisfy oneself that the tasks and items have been well chosen, and decide for oneself to what extent he agrees that the items are really relevant to an elementary knowledge of French.

While questions with uncued responses may be typical of natural conversational exchanges, picture-cued responses are less so, and the question-formation tasks appear to be totally unnatural. A certain validity must of course be granted such criticisms. On the other hand, one may note that the picture-cued responses make fewer demands on student recall and limit the range of possible answers to a greater degree than the uncued responses. At the same time, they provide a context for testing a wider range of vocabulary items. With regard to the question formulation section, one may observe that it is difficult to imagine *any* non-artificial means to evoke interrogative structures, yet such structures are an integral part of communicative ability and should therefore be tested. One might simply view this part of the test as a simple reversal of the typical conversational sequence; thus, it is not totally alien to communicative competence.

Like many objective-type tests, our test is inevitably prey to the criticism that the test conditions—the laboratory setting, microphone, and loud speaker—render it an invalid measure of students' ability to communicate in a natural conversational setting. Indeed, an oral interview comes closer to being valid in that regard. However, it seems that very few students visibly allow the unnatural setting to shake their ability to respond (as also supported in the congruent statistics).

3.3 *Reliability*

The term "reliability" refers to the dependability of the test—is it a consistent measuring device? We may speak of two types of reliability: (1) reliability of the test, i.e., is the test itself constructed in such a way as to yield consistent results? (2) reliability of scoring: is consistency of results facilitated by the nature of the scoring process?

3.31 *Reliability of the test*

In determining the reliability of the test, various factors must be considered. Rebecca Valette[4] cites the requisites of a dependable test as standard tasks, standard conditions, standard scoring, and multiple samples. The test satisfies all four criteria. All students take the same test, thus perform the same tasks. The testing conditions are the same insofar as possible, the only variable being the physical and psychological state of the student, which is beyond the control of the examiners. Scoring is standardized, as described earlier, to the extent such standardization is feasible.

An objective measure of adequacy of sampling can be obtained by statistical procedures such as item analysis and split-half correlation. Though neither

[4] Rebecca M. Valette, *Modern Language Testing: A Handbook* (New York: Harcourt Brace, Jovanovich, 1967), p. 31.

of these procedures has been applied in this case, we have concluded, by simple inspection of the test and a sample of student responses, that the items included are of a sufficiently wide range of difficulty. There is a gradation in the level of difficulty from certain "easy" items on which most students tend to score high, such as *Quel âge a ton père?* or items that may be answered by partially memorized responses (*Quelle heure est-il?*) to questions involving rather complex grammatical structures and requiring the perception of non-redundant grammatical cues, as in the question *Où est-ce que vous allez pour étudier le soir?* A statistical index of the test's reliability was obtained by analyzing the results of 138 tests administered at the end of the academic year 1976–1977. The tests showed a range of 36–100, a mean of 75.5, and a median of 78–79. The standard deviation was 13 and the standard error 1. Using the Kuder-Richardson Formula 21, the reliability of the test was calculated to be .87, which is quite high for "homemade" tests and approaches the .90 required for standardized tests (although many so-called standardized tests fail to meet this reliability quotient).[5]

3.32 *Reliability of scoring*

Any test that departs from the discrete-item format presents problems of intersubjective and intrasubjective reliability. Both types of reliability may be increased by a detailed description of scoring criteria of the sort we have attempted in 2. above. Intersubjective reliability may be enhanced by providing scorers with extensive training and the opportunity to accommodate their judgments. In order to assess the maximum intersubjective reliability of the IUFCAT two associates of the authors of this article (B. Kerr-Barnes and B. Tice), who at the time had served as graduate associate instructors in the Indiana University beginning French program for two years and had considerable experience with the instrument, evaluated independently a set of twenty test tapes. Before scoring these tapes, they engaged in limited discussion about scoring criteria. It should be pointed out that their assessment involved an earlier version of the IUFCAT in which the appropriateness and well-formedness scales of part I and part II comprised only three marks (0–1–2), instead of the current four marks, and the fluency and pronunciation accuracy contained only two marks (0–1). The resultant scores (see Table 1) were compared by means of a rank-order correlation and the comparison yielded a correlation coefficient of .94.

This high correlation coefficient establishes the ability of the scoring scheme to provide for the accurate ranking of student performances relative to each other on the part of different scorers. However, it is evident that the two scorers differed, sometimes significantly, in the total scores awarded individual students. Except in four cases, the difference is always in the same direction, which indicates it is due to a difference of severity between the scorers.

[5] Valette notes that the reliability of a good classroom test falls between .60 and .80.

TABLE 1

Statistical Analysis of Intersubjective Reliability

Student	PARTS I AND II		PART III		TOTAL		Difference in total
	Scorer 1	Scorer 2	Scorer 1	Scorer 2	Scorer 1	Scorer 2	
A	22	22	14	15	36	37	+1
B	25	26	4	6	29	32	+3
C	19	19	1	2	20	21	+1
D	38	31	31	31	69	62	−7
E	36	35	32	33	68	68	0
F	38	38	32	32	70	70	0
G	32	35	24	24	56	59	+3
H	42	42	29	29	71	71	0
I	33	37	20	24	53	61	+8
J	40	42	35	34	75	76	+1
K	38	38	28	26	66	64	−2
L	24	27	4	8	28	35	+7
M	21	17	4	6	25	23	−2
N	27	28	31	30	58	58	0
O	11	15	2	4	13	19	+6
P	31	31	24	25	56	57	+1
Q	38	38	25	28	63	66	+3
R	16	23	18	17	34	40	+6
S	26	33	24	26	50	59	+9
T	34	30	19	19	53	49	−4

Rank-order correlation coefficients

Part II = .91
Part III = .99
Total = .94

Separate correlation coefficients of .91 and .99 were found for parts I and II versus part III, respectively. This difference was attributed to the structure of these sections of the examination; part III generated responses that showed less variety because of the nature of the task required. Another contributing factor appeared to be the difference in the scoring scales for these sections. In order to improve scoring reliability, it was decided to use the same scale for parts I, II, and III.

4. Conclusion

Despite inherent shortcomings in the area of scoring reliability, it seems to us that the IUFCAT has proven to be a generally valid, reliable, and, above all, practical, measure of minimal communicative ability in French. Because it is standardized, it is better suited for administration to large groups than an oral interview and, with respect to this alternative approach, requires evaluators with considerably less experience and training. As compared to the two existing tests of speaking proficiency (MLA Cooperative and Pimsleur tests), it shows superior validity since it is a more direct measure of

communicative ability. The structure of its rating scales also appears to give it greater scoring reliability. We believe, therefore, that this test provides a worthwhile model for the elaboration of instruments measuring communicative ability suitable for use in multi-section FL programs where test samples are large and where teaching personnel are likely to be relatively inexperienced.

In concluding, we would like to mention some other advantages that may be gained from the use of formal measures of communicative ability such as the IUFCAT.

The IUFCAT provides a representative sample of students' range of communicative ability. This enables us to formulate realistic course objectives in that skill and facilitates course planning and syllabus design. The test also serves as a valuable training tool for new instructors. By providing concrete samples of course-final performance, it defines a point of reference against which instructors can measure student progress during the course. Thus it serves as a guide for correction of errors and for the planning of classroom activities that will lead to the level of communicative ability exemplified by course-final student performance. The test also defines for students the level of accomplishment expected of them.

From a more research-oriented perspective, the IUFCAT generates abundant samples of learner interlanguage. The analysis of inappropriate and ungrammatical utterances may lead to new insights on the process of language acquisition and on learning strategies. These in turn may provide firmer basis for teaching procedures, for the design of syllabi, and for the preparation of more effective pedagogical materials.[6]

INDIANA UNIVERSITY

[6] Scoring sheets for the IUFCAT and a tape containing test items and sample student responses are available for general distribution at the cost of $5.00 (tape or cassette); address requests to A. Valdman, Department of French and Italian, Indiana University, Bloomington, IN 47405.