
Language Testing: Past and Current Status - Directions for the Future

JOHN L. D. CLARK

THE TWO MAJOR PURPOSES OF THIS PAPER ARE: 1) to characterize the historical and present situation in foreign and second language testing in the United States with respect to the major measurement trends represented; and 2) to suggest a number of desirable development activities over the near- and mid-term future that may help to build, in an evolutionary way, on several current initiatives in the field—initiatives that already show, as of the early 1980s, substantial promise for facilitating and reflecting effective, performance-based language instruction to an extent not heretofore reached or even closely approximated.

PREVIOUS TRENDS

In order adequately to discuss the current situation in foreign/second language testing, it will be helpful to review briefly the history of development of language testing theory and practice over the last several decades. In 1978, Spolsky identified three major historical trends in language testing, which he referred to as “pre-scientific” (roughly prior to the early 1950s), “psychometric structuralist” (early 1950s through about the late 1960s) and “integrative-sociolinguistic” (late 1960s and following).¹ During the “pre-scientific” period, there was, in general, little concern about or attention paid to the reliability, validity, or other important psychometric characteristics of the testing activities carried out in the course of language instruction. In keeping with the grammar-translation, reading-oriented teaching approaches widely used at that time, classroom instructors found it both acceptable and adequate for their purposes to present the students fairly lengthy passages for translation from or into the target language, exercises on selected grammatical

points (often dealing with special usages or exceptions to the general rules), and various cultural items, predominantly of the “capital C” type. Teachers acted autonomously in preparing and grading their own tests, and it was implicitly assumed that any individual who was qualified to teach a language was, by the same token, fully competent to handle its assessment. Although the development of functional proficiency in listening comprehension and speaking-or, for that matter, of reading and writing competence in other than relatively artificial, academic contexts—was not at issue during the “pre-scientific” period, the testing approaches utilized during this period would obviously not for the most part have served to measure validly the student’s ability to make use of the language as a means of functional communication in “real-life” situations outside of the classroom setting.

The second period, the “psychometric-structuralist,” may be considered to have evolved in very large part from the theoretical work and related empirical studies carried out by Lado in the mid-fifties.² The basic orientation for Lado’s approach to testing grew out of, and closely paralleled, the then current structural-analytic approach to linguistic research, as exemplified, from a language teaching standpoint, in a variety of contrastive analysis studies comparing selected target languages to English with regard to both phonology and structure.³ Adopting a fundamental pedagogical assumption underlying the contrastive analysis approach—that it was necessary formally to teach only those, particular features of the target language that were predicted, through contrastive analysis, to pose learning problems for the student—Lado extended this assumption into the testing area by asserting that measurement of the student’s language performance could most efficiently and effectively be accomplished by limiting the testing activity to the same points of

difficulty. As explicitly stated by Lado: "Since some (aspects of the target language) are easy to master because they are already known from previous language training in mastering the native language, we will generally eliminate these from the corpus. We will attempt to test the learning problems, on the grounds that knowing the problems is knowing the language. We say specifically that testing the problems is testing the language."⁴

To determine, to as unambiguous a way as possible, the student's command or lack of command of each of the various "problem" areas identified by prior contrastive analysis. Testing formats were required that were much more highly focused than the open-ended and essentially adventitious passages and exercises of the pre-scientific period. The bulk of Lado's 1961 text was accordingly devoted to presenting and explicating a number of carefully developed formats for testing a variety of specific language elements such as the aural discrimination of phonemes (e.g., "sheep" vs. "ship" for Spanish speaking learners of English); recognitional or productive control of specified lexical items; perception of syntax-mediated meaning differences in otherwise identical utterances (e.g., "The boy hit the car" vs. "The car hit the boy"), and so forth. In all instances, language aspects other than those comprising the particular point being tested were either eliminated from the test altogether or kept to the bare minimum needed to frame the tested point linguistically. The one-tested-element-per-test-item procedure followed by Lado and, subsequently, by a number of other test developers was referred to by Carroll in a 1961 article as the "discrete-point" approach -a term generally associated with this type of testing since that time.⁵

From the perspective of effective measurement of functional language competence, the approach developed by Lado represented some degree of advancement over earlier testing practices in that it explicitly addressed the student's ability or lack of ability to perform each of a number of specified linguistic tasks in the target language. By its very nature, however, the discrete-point, one-element-per-item testing procedure was not capable of measuring the student's ability to comprehend or produce, on a holistic basis, a larger and more natural corpus of language material than that

represented material than that represented by individual-element test questions.

The test-related work carried out by Brooks over roughly the same time period as that of Lado has also generally to keeping with the discrete-point orientation. However, Brooks' approach was considerably less dogmatic, as witnessed in the general format of the NDEA supported *MLA-Cooperative Classroom Achievement Tests* in French, German, Italian, Spanish, and Russian, for which Brooks served as consultant.⁶ Although these tests made use of a number of discrete-point item types (for example, the verbatim repetition of short spoken utterances, scored for accuracy of pronunciation of specified phonemes), they also included a number of more naturalistic, "real-life" exercises, such as the presentation of reasonably lengthy tape recorded dialogues between two native speakers, followed by questions covering general comprehension.

The tentative, partial expansion beyond discrete-point oriented techniques noted in the *MLA-Cooperative tests* (as well as in the *MLA Proficiency Tests for Teachers and Advanced students* published at about the same time) developed into almost wholesale rejection of the discrete-point approach during the period referred to by Spolsky as "integrative-sociolinguistic" (late 1960s and following).⁷ During this period, the theoretical orientation adopted with respect to testing reflected in large part a growing dissatisfaction with structural linguistic theory as a proper model for language analysis and, by the same token, pedagogical practice. Contemporary linguistic research held that the use of language for real-life communication involved a creative act in which the whole of the communicative event was considerably greater than the sum of its linguistic elements. As a result, the adequacy or effectiveness of the communication could not be adequately assessed through individual evaluation of its component parts.

Assessment procedures deriving from the new linguistic orientation, and which undertook to determine the student's ability to carry out more globally-oriented language, use tasks extending beyond individual-element performance, came to be referred to as "integrative" (a term also derived from Carroll). Oller was among the first and most ardent proponents of integratively-oriented testing procedures. Over a period of several years beginning about 1970, he and several of his graduate students

and other associates reported a number of detailed experimental studies on the "cloze" technique - a procedure originally developed by Taylor in 1953 to measure reading proficiency to English as a native language and in which the student is required to resupply individual words systematically deleted from a continuous printed text.⁸

Cloze-based studies by the "Oller group" and others included analyses of the effects on reliability and validity of various cloze scoring procedures;⁹ examination of the measurement consequences of restricting the deleted lexical items to a single linguistic category (e.g., prepositions) rather than selecting them on an "every *n*th-word" basis;¹⁰ theoretical and practical considerations associated with the development of a multiple-choice format for cloze;¹¹ and the measurement appropriateness and practicality of testing listening comprehension through a "spoken cloze" technique.¹² Two extensive bibliographies on cloze testing compiled by Oller¹³ and by Reilly¹⁴ provide a good indication of the extent of measurement interest generated by the cloze technique.

In addition to the doze procedure, Oller and others experimented with several other integratively-based testing techniques, including the traditional dictation¹⁵ as well as a "reduced redundancy" procedure in which the examinee is asked to carry out aural discrimination tasks under conditions in which the spoken material is intentionally degraded by the introduction of various amounts of white noise.¹⁶

The emphasis accorded during this "integrative" period to the presentation of continuous, natural language texts from which the student is asked to derive elements of meaning that go beyond single-sentence boundaries may be considered to approach the testing of functional language in a communicative context even more closely than was the case with discrete-point tests of the "psychometric-structuralist" type. However, dictation, doze procedure, and reduced redundancy testing all fall somewhat short of representing genuine communicative activities in that the tasks performed by the student in working with these tests would only rarely be at issue in real-life language-use settings. Except for such relatively

infrequent situations as taking notes over the phone, reconstructing a poorly handwritten text or garbled telegram, or understanding speech over a defective telephone connection, the linguistic situations presented and student performances elicited in dictation, doze, and reduced redundancy testing are not part of everyday language use and cannot be considered face- and content-valid representations of the kinds of language situations and performances at issue in the mainstream of real-life language use.

CURRENT TRENDS

Since the publication of the Spolsky article, increasingly wide interest has been shown and a variety of developmental activities carried out in an area frequently referred to as *direct proficiency testing*, involving the design and use of assessment techniques which require the examinee to perform functionally oriented language use tasks in situations that approximate as closely as possible the conditions under which these tasks are carried out in the real-life setting. Chief among the current procedures aimed at direct proficiency measurement is the Foreign Service Institute (FSI) oral interview, a testing technique that has recently not only received considerable theoretical and research attention from language testing specialists but also become a matter of considerable practical interest to front-line language teachers and administrators.¹⁷

Although widespread popular interest in FSI-type interview testing has arisen only recently, the original development of this testing technique and the associated rating scale dates back to the early 1950s-contemporaneously with the much more widely publicized discrete-point approaches then being developed within the psychometric-structuralist framework. As detailed in its historical perspective by Wilds¹⁸ and by Sollenberger,¹⁹ the "FSI interview" was originally developed within that agency as a means of determining the extent to which graduates of the language training program would be able to function in a linguistically appropriate and effective manner in the particular language-use situations they would be expected to encounter in carrying out their various assignments abroad. In recognition of the fact that the majority of these situations would involve face-to-face discussions with native speakers, it was decided to make use of direct conversation with a native speaker as the basic

testing procedure. Further, in order to probe fully the student's ability to communicate effectively on a variety of topics and within a number of different contexts, the interview conversation was carefully structured so as to present, within the available testing time, numerous communicative situations, beginning with polite social conversation and progressing through increasingly challenging and sophisticated linguistic tasks, up to the point of maximum possible performance by the examinee.

In keeping with the wide range of proficiency covered in the test, a correspondingly broad scoring scale was developed, ranging across six verbally-defined levels from "no functional proficiency" in the target language (level 0) to proficiency indistinguishable in all respects from that of an educated native speaker (level 5). The verbal description of level 2 on the official FSI scale gives a general indication of the nature of the level descriptions and the degree of detail which they provide:

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

In addition to the six verbally defined levels, "plus" values were subsequently added to each level (except level 5), to be awarded when the examinee "substantially exceeded" the performance required at a given level but failed to meet in all respects the requirements of the next higher level. Including "plus" values, the total FSI scale recognized a total of eleven levels.

Following its development, and up until the early 1970s, use of the FSI interview technique is largely concentrated within the Foreign Service Institute and certain other government agencies. However, during the early and mid-seventies, use of, and general familiarity with, the FSI interview increased substantially, largely as a result of a

series of testing projects carried out on behalf of several client groups by the Educational Testing Service. Through successive contracts with the Peace Corps beginning in 1969, ETS language staff conducted numerous interviewer training workshops for Peace Corps in-country staff and, as of late 1977, some 560 testers in fifty-five countries had been trained in the FSI technique and had administered a total of approximately 18,000 interviews. Subsequent ETS-coordinated tester training sessions for secondary school teachers of French and English as a second language in New Brunswick (Canada)-as well as the introduction of a New Jersey state teacher certification program for Spanish-English bilingual and ESL teachers which involved, among other requirements, obtaining a satisfactory rating on an FSI-type interview-contributed substantially to increased general awareness within the language community of the interview technique and its potential applications beyond the original government service context. Some indication of the amount of both popular and research-oriented interest in the FSI interview that had been generated by the late 1970s may be seen in the proceedings of a two-day conference on direct proficiency testing conducted in March 1978, attended by over 200 participants and involving the presentation of eighteen technical and practically-oriented papers on interview-related topics.²⁰

Further public exposure was given the interview-based testing approach by the FSI itself, which, during 1979-80, convened a series of invitational conferences of college foreign language department chairmen, state foreign language supervisors, curriculum developers, and others to familiarize the participants with the nature of and rationale for interview-based proficiency testing and to explore the usefulness of this technique in typical academic settings.²¹

To further disseminate information about and provide hands-on experience with the FSI technique, in February, 1982, the American Council on the Teaching of Foreign Languages (ACTFL), working with ETS and with funding support by the Department of Education, conducted the first of a planned series of intensive interviewer, rater training workshops for language teachers at the postsecondary level. These workshops have been designed to train the participants to the point at which they are able to administer and score FSI-type

interviews in the same manner and with the same degree of accuracy as FSI testing staff and others formally certified in the testing process. Less intensive “familiarization” workshops, as well as more comprehensive tester training sessions, are currently being offered by ETS on request from undergraduate language departments, secondary school systems, or other interested groups.

From the viewpoint of its potential impact on foreign/second language teaching practice, widespread utilization of the FSI-type interview (or of appropriate and carefully planned modifications of it) holds enormous promise for substantially advancing the development of functionally-oriented language training programs in the United States. Equipped with an appropriate, external-to-program indicator of acquired proficiency in the target language, referenced against the performance requirements inherent in real-life language use, it becomes possible for a variety of individuals and groups—ranging from classroom teachers and their students, through local schools and school systems, to planners and implementers of broad-scale studies of the “national yield” of current language training—to determine the functional outcomes of this instruction and to suggest possible further improvements in the instructional process on the basis of the information obtained.

A related trend, which shows substantial promise of encouraging and facilitating language instruction aimed at functional proficiency development, was expressed in broad outline form by Omaggio in an article prepared for the 1980 National Conference on Professional Priorities.²² The significant conceptual contribution made by Omaggio in this article was to relate teacher-prepared classroom testing activities to functional proficiency development concerns by suggesting that, notwithstanding the fact that the linguistic material that can be effectively dealt with in typical academic settings is necessarily restricted in terms of both the range of grammatical features and lexical items represented, the classroom teacher should nonetheless make every effort to teach and test this reduced corpus within the context of realistic, functional language-use situations having genuine communicative relevance for the student. Toward this end, Omaggio described and provided examples of a number of what she termed “hybrid”

tests which “artfully combine grammar and context, structure, and situation” into testing exercises that, while limited to the specific structures and general areas of lexicon dealt with in the course, reflect real-life communication situations to the greatest possible extent. For example, as a test of lexical control, an exercise in which the student is asked to write a letter to a real or imagined “pen pal” on a topic for which the relevant vocabulary has already been introduced in class would be considered a much more desirable measurement approach than, for example, requiring the student to write out a series of unrelated sentences involving the same vocabulary items. Omaggio gave the following guidelines for developing testing activities incorporating the “hybrid” principle:

- 1) the situation depicted is relevant and immediately useful to the language learner;
- 2) the content reflects the level of sophistication of the students and their knowledge of the world;
- 3) the language is, at all times, natural, respecting the ‘conditions of elicitation’ of certain types of structures in natural language use;
- 4) answers required of students have ‘truth value’ (that is, do not involve imaginary or contrary-to-fact responses from the student’s own perspective);
- 5) characters used in (test) items are ‘realistic,’ in that they have personality and relate to the learner’s experience in some way;
- 6) items respect sociolinguistic norms;
- 7) the language sample is short enough so that students have little difficulty remembering it, but long enough to provide the necessary context.²³

Three major advantages of this communicative orientation to classroom testing may be identified. First, on the assumption that the ultimate objective of the language teaching process is to develop as fully as possible the student’s capacity to use the target language appropriately and effectively in genuine communication settings, classroom exercises and associated testing activities that approximate these settings as closely as possible would more directly, and presumably more effectively, address this goal than would less highly proficiency-based teaching and assessment procedures.

Second, communicative testing procedures in all probability correspond more closely to teachers’ natural tendencies on assessment matters than do

either the highly formalized, discrete-point testing approaches of the psychometric-structuralist period or the cloze tests and other procedures typical of the integrative-sociolinguistic orientation. To the extent that classroom teachers are able readily to understand and personally relate to the types of testing exercises at issue in Omaggio's proficiency-oriented approach, their willingness to expend the time and effort required to develop and make effective use of these procedures will in all probability be increased beyond the relatively modest levels of interest and implementation accorded earlier prescriptions for classroom testing.

A third consideration favoring increased attention to and broader dissemination of proficiency-oriented classroom testing procedures is the high level of student interest and motivation which these procedures appear to engender by comparison to less realistic testing techniques. For example, in a recent study, Brutch found that her students exhibited significantly more positive attitudes toward taking a communicatively-oriented test involving the writing of essays and personal letters than they did toward working on the writing section of the more highly discrete-point *MLA-Cooperative French Proficiency Test*.²⁴ Schulz reported an experiment in which eighty college students were administered both a series of discrete-point tests and several "simulated communication" tests such as following map directions on the basis of instructions given in the target language. The great majority (78.8%) indicated that they preferred taking the simulated communication tests.²⁵ Although student opinion might be accorded somewhat lower priority in classroom testing decisions than a number of other considerations, testing approaches that give rise to generally positive reactions on the students' part may be considered preferable to those that are routinely received with hostility or, at best, indifference.

An important caveat concerning the classroom testing approach advocated by Omaggio must be introduced at this point. Although the kinds of testing exercises which she recommends may quite closely approximate both the general testing formats and language tasks associated with proficiency testing activities, they cannot properly

be considered true proficiency tests in that their content is, intentionally, limited to those linguistic aspects that have been dealt with in the course of instruction. Unless students, teachers, and others associated with the administration and interpretation of proficiency-oriented classroom tests keep this consideration closely in mind, and unless there is periodic administration - for cross-reference and comparison purposes - of a genuine proficiency criterion such as the FSI-type interview, the distinct possibility exists that the classroom testing results will in fairly short order come to be generously interpreted - that is, viewed as reflecting a higher level of linguistic competence than would, in fact, be indicated by administration of an appropriate external proficiency test.

A third current trend in the language teaching field - a trend that has not yet been explicitly related to language testing to any appreciable extent but that shows considerable potential for becoming an important and integral component of the overall "measurement operation" - is the use of computers in the service of language instruction. Within the past two to three years, a large number of journal articles (see Meredith's essay elsewhere in this issue), conference presentations, local and regional workshops, and other media and activities have been concentrated in this area, which shows prospects of exponential growth within the immediate future. In a survey of colleges and universities conducted during the 1978-79 academic year, Olsen found sixty-two institutions at which some form of computer-assisted language instruction (CALI) was taking place; fourteen additional institutions reported that they intended to introduce such a program within the next two years.²⁶ Reports of CALI activities in the recent literature include descriptions of "mainframe" language instruction on the PLATO system at the University of Illinois-Urbana,²⁷ computerized drills on vocabulary and grammar at the University of Virginia,²⁸ and at MIT,²⁹ and a computer-based supplementary grammar program in German at Ohio State University.³⁰ At the primary and secondary school level, although there are few detailed reports of computer use in second language instruction, there is little doubt that the necessary hardware is being rapidly put into place. The most recent report of the National Center for Education Statistics shows a threefold increase over the fall, 1980, figures in the

number of microcomputers in use in public schools: an estimated 4.7 million student made use of computers in some instructional capacity during 1981-82.³¹

With respect to the development and dissemination of computer programs for language instruction, at the November 1982 ACTFL Annual Meeting, software packages were prominently displayed at a number of exhibit booths, including those of several major publishers. In a review article for the February 1983 *Northeast Conference Newsletter*, Harrison listed a total of 103 commercially available microcomputer programs for language instruction, ranging in scope from simple structural or vocabulary drills to considerably more extensive learning activities.

³² Given both the widespread availability of appropriate computer hardware and rapidly growing interest on the part of both language professionals and commercial publishers in developing instructional materials that take advantage of this technology, it would appear that a very important priority for the language teaching field would be to carry out the conceptual development and related empirical studies required to determine the respective roles to be played by the live instructor and by the computer within a total instructional process that maximizes the instructional capabilities of both.

I would suggest that by far the most appropriate assessment-related application of computer capabilities (at least over the next several years pending major advances in voice recognition and speech synthesis) will be in the diagnostic testing, on a real-time basis and with either immediate or very rapid feedback to the student, of those language elements which are most amenable to effective instruction through computer-assisted means, and including, for example, vocabulary development, initial instruction and drilling on morphological and syntactical features, and a variety of types of exercises involving the training of reading comprehension. Except for this last category, which involves somewhat more global types of language use, all of these areas of "computer strength" (insofar as its instructional capabilities are concerned) are also areas that call for highly diagnostic, element-by-element assessment procedures.

As discussed elsewhere in greater detail, an enormous amount of carefully planned and systematic effort is required to develop and administer a classroom diagnostic testing program of even modest scope.³³ For example, the development and use, over a school term, of ten 40-item diagnostic tests with a class of thirty students would involve the initial preparation of 400 different questions. This would be followed by a series of administrations resulting in the generation, in total, of 12,000 separate items of information about the students' performance: each of these items would, in turn, require initial evaluation (scoring) followed, presumably, by a variety of tabulations intended to determine not only strengths and weaknesses in the performance of particular students (for targeted remediation on an individual-student basis) but also group performance profiles for given language elements (as a guide to possible needed revisions in the teaching process for particular elements). Given the sheer logistic and time-utilization problems involved, it is not surprising that even the most dedicated teachers would hesitate to undertake a thoroughgoing diagnostic testing effort if they were the only "resource" available for this purpose.

Although the information-handling implications of diagnostic testing are of a magnitude to give considerable pause to the live instructor, the error-free and virtually instantaneous accomplishment of these same tasks is a trivial operation for mainframe computers and is well within the technical capabilities of virtually any microcomputer that would seriously be considered for purchase and use in an academic setting. Even more impressively, beyond the simple two-way tabulation of fixed data items involved in the preceding example (with all students working on all test questions in a uniform, sequential fashion), current microcomputers are, for the most part (and assuming appropriate programming), sufficiently powerful to select particular test questions to be presented to a given student on the basis of the student's responses earlier in the testing sequence). For example, the computer may be programmed so that students who answer correctly each of three separate items on a given verb form are automatically credited with "knowing" that form, and no further items on that particular element are presented. On the other hand, students who show some weakness in a particular area can be presented increasingly detailed questions to

determine the exact parameters of weakness involved. (In this example, the student might be posed a series of additional questions probing that particular tense with respect to first-person forms, second-person forms, etc.).

FUTURE DIRECTIONS

The continued development and synergistic interaction of the three major trends identified above-increased interest in and utilization of external-to-program, "real-life" oriented measures of functional language proficiency; an emphasis on classroom testing that stresses functional language use within the context of the particular learning experiences of the students as of the time of testing; and more widespread utilization of computer capabilities for both instructional and testing purposes - will be such as to create an unprecedented opportunity for the assessment endeavor to more closely parallel and more effectively serve the cause of proficiency-oriented language instruction than has ever been the case in the past. Whether or not these trends will, in fact, continue and develop fully over the near- and midterm cannot be determined with confidence at the present time; instead, the discussion below is intended to suggest what might be considered the major components of a "desirable future" outcome in these areas.

Increased research in and practical attention to curriculum-free direct proficiency testing as the fundamental performance benchmark for both individual students and language training programs. A number of research studies have already been undertaken with respect to major psychometric aspects of the FSI-type interviewing and rating procedure, including inter- and intra-rater reliability studies;³⁵ concurrent validation with other instruments such as the *MLA-Cooperative Foreign Language Proficiency Test*:³⁵ construct validation using the Campbell and Fiske multitrait-multimethod procedure;³⁶ investigation of the functional relationships between the "global" proficiency rating and sub-ratings of "pronunciation," "listening comprehension," "vocabulary," "structure," and "fluency";³⁷ and the degree of correspondence between interview level scores and student self-ratings of speaking proficiency.³⁸ Although these investigations have, in general, been well conducted and have provided a large amount of information

against which the over all validity and applicability of the direct interview technique can be evaluated (with generally positive results), there are a number of other aspects of both the interview technique and scoring procedure that will require additional research and developmental attention if the assessment promise of this type of testing is to be fully realized.

A major concern is that the sociolinguistic context of the interview-polite formal conversation with a relative stranger-represents only one of a variety of communicative roles that the examinee would be expected to play in language-use situations outside of the classroom setting. Speaking situations in which, for example, other than neutral effect (e.g., anger, doubt, dissatisfaction) must be conveyed; which involve unequal status between the interlocutors (e.g., employer-employee relationships); or which take place under less than optimal acoustic conditions (e.g., in a noisy restaurant), are not well represented in the interview in its usual configuration. Some attempts have been made to incorporate more diverse language use settings by having the examinee, in the course of the interview, carry out one or more role-play tasks, typically cued by printed native language instructions spelling out in considerable detail not only the communicative task involved but also the major sociolinguistic variables at issue in the situation. An example of one such description is as follows: "You have made an appointment to have lunch at noon on the next Monday with an older business acquaintance whom you do not know well. On that day, you are very busy and completely forget about the date until 12:30. You hurriedly find a taxi and rush to the restaurant to find your luncheon partner waiting outside for you since they would not let him inside until you arrived. What would you say?"³⁹

Although role-play activities such as these certainly do not parallel genuine language-use situations in all communication-relevant respects, they do make some provision for introducing into the testing setting a variety of language styles. Registers, and affective elements that would not otherwise be probed; as such, this technique would be considered to warrant detailed and systematic development within the near term.

Directly related to the problem of expanding the testing procedure to include a variety of language

usage beyond that involved in "police conversation" is the matter of specifying scoring procedures that take adequate account of the sociolinguistic aspects at issue. In 1980, Canale and Swain proposed a "model of communicative performance" which included, along with the usual grammatical and lexical components, the two additional categories of "sociolinguistic competence" and "strategic competence."⁴⁰ These two terms referred, respectively, to the speaker's ability to make use of the appropriate register and tone for a given interlocutor and communicative situation: and to properly "manage" the conversation by, for example, adhering to proper turn-taking conventions, making appropriate use of gestures, asking for and providing clarification of meaning where necessary, and so forth. Some early efforts to incorporate such elements into the interview rating process have been made by Ingram and Wylie in their *Australian Second Language Proficiency Ratings scale*, in which, for example, the speaker at "level 4" is required to have "considerable sensitivity to register requirements" and to be able to "(readily modify) the language appropriately."⁴¹

The Foreign Service Institute has itself recently experimented with both sociolinguistically-oriented changes in the interviewing procedure and corresponding modifications in the rating scale, including the incorporation of a "discourse competence" factor involving both conversational management strategies and the ability to sequence appropriately and combine elements of information within fairly lengthy utterances. With respect to potential future modifications in the interview rating scale, it will be highly important to insure that these modifications are based on the best available theoretical and empirical information concerning the influence on communication of the sociolinguistic variables involved.

It will be equally important to determine that the rating procedures finally developed call for evaluative judgments that can, in fact, be made in a straightforward and reliable manner by those individuals who will be involved in the rating process. At present, the probable level of scoring reliability for such aspects as "proper use of nonverbal cues" or "selection of appropriate register" is not generally known, and although such judgments might be consistently possible for persons having a high level

Of linguistic background and training, it is unknown whether such an ability would be intrinsically present in the typical classroom language teacher or could be adequately developed within a tester training program of reasonable cost and duration. Wider-scale dissemination and effective use, at the local school or college level, of interviewing/rating procedures incorporating these or other "sociolinguistic elements" would be expected to depend in large part on a positive response to this question.

A third major conceptual and practical problem that arises in connection with use of the FSI-type interview in academic contexts is the wide range of proficiency covered in the test (no useful proficiency in the language up to that of an educated native speaker). General familiarity with the proficiency outcomes of typical classroom language programs-as well as the results of an earlier nationwide study by Carroll in which the average FSI score of graduating college language majors was found to be only slightly beyond level II - would suggest that the anticipated level of proficiency for students completing secondary school or undergraduate programs would, in general, be limited to the lower portion of the FSI scale.⁴² The negative motivational implications, for both students and instructors, of having students show little or no progress on the regular FSI scale even after a reasonably extended period of instruction cannot be taken lightly in considering the academic use of this testing approach.

In an effort to resolve some of the problems associated with the coarseness of the traditional FSI score intervals, ACTFL has recently received funding support for and coordinated the development of a series of "expanded" proficiency descriptions at the lower end of the FSI scale (see Omaggio's essay elsewhere in this issue). Working titles for these descriptions (Novice Low, Novice Mid, Novice High, Intermediate Low, Intermediate Mid, and Intermediate Advanced) were chosen to express more positive connotations about the student's language attainments than the "survival" and "limited working proficiency" designations of the original FSI levels one and two.⁴³ A basic question in regard to the use of these expanded scales as meaningful indices of language proficiency is whether the verbal distinctions drawn between these levels do in fact reflect meaningful differences, with respect to the

types of language-use tasks which the examinee would be able to carry out in real-life communicative situations outside of the classroom setting. If, for example, both Novice Low and Novice Mid students fail to reach the baseline level of "linguistic survival" represented by level I on the regular FSI scale, there would seem to be little justification in considering that they differ in any meaningful way with respect to overall language proficiency.

A second, related concern is whether practicing teachers and other users of the expanded scale will in fact be capable of making, in a reliable way, the rather fine-grained distinctions represented by adjacent score levels. In the absence of reasonably high scoring reliability among teachers and other in-field users of the scale, there is the possibility that, for example, a number of "Novice Mid" students entering a language class in September would be scored as "Novice Low" the following June. The negative pedagogical and public relations implications of such a testing outcome are fairly obvious, as is the need for focused research attention to the measurement considerations involved.

The direct testing of speaking proficiency, as exemplified in the FSI interview and other modifications of this general approach, shows excellent prospects of continuing to be a major area of interest for both testing specialists and the language teaching profession over the next several years. A number of both theoretical and practical questions must still be addressed in taking the interview technique from its original development and use context into broader areas of application, and the longer-term viability of this approach will depend crucially on the extent to which these questions are forthrightly addressed and satisfactorily resolved.

Further development of diagnostic testing principles and procedures as an integral aspect of computer-assisted instruction. Assuming that instruction in such relatively rote aspects of language learning as vocabulary development and structural explanation and practice will increasingly become the province of the computer, the corresponding measurement role for this technology will be the ongoing diagnostic evaluation of student performance in these particular areas, with constant relevant feedback provided to both student and teacher. By comparison to other

areas of language testing, the development of diagnostic testing and reporting procedures (for either human or computer-assisted implementation) has up until now received relatively little attention. For example, a December 1982 search of the ERIC (Educational Resources Information Center) database yielded only seven citations combining "language testing" and "diagnostic testing" descriptors. Jones is currently conducting a project funded by the Department of Education to investigate the use of the microcomputer for diagnostic testing, but results of this study are not yet available.⁴⁴ Research and development matters that need to be addressed in this area include; 1) specification of question formats that unambiguously focus on the testing point at issue; 2) maintenance, to the greatest extent possible, of realistic linguistic contexts for the test questions; and 3) development of tabulation procedures and student/instructor feedback protocols that are immediately meaningful to the recipients. (With respect to individual students, diagnostically-oriented feedback would presumably include specific reference to appropriate instructional material for all items not yet mastered, with immediate "at-the-terminal" access to such material.)

Additional complexities in, as well as yet un-envisioned advantages of, computer-assisted instruction and diagnostic testing in those aspects of the language teaching process for which the computer is particularly well suited will in all probability be identified in the future; in any event, major conceptual and practical advances in this area will take place as a direct function of the extent to which computer technology is incorporated into the mainstream of language instruction as an integral part of the overall instructional system.

Modification of the teaching/assessment role of the teacher toward explicit proficiency goals. Although it is not possible within the scope of this paper to detail the variety of initiatives that are being taken by ACTFL and a number of other organizations to encourage and provide support for proficiency-oriented language teaching, it is possible to identify the broad components of an optimum *measurement role* for the classroom teacher within this instructional context. I would suggest that this role will involve three major elements; 1) a working familiarity with the rationale for and techniques involved in direct

direct proficiency interviewing and related testing procedures, 2) the ability readily to prepare and use proficiency-oriented achievement tests based on material covered in the course; and 3) the ability to use and interpret properly the measurement-related data provided by computer-based teaching and testing procedures supplementing the live classroom instruction.

With respect to the first element, it will become increasingly important for teachers, supervisors, and others involved to the language teaching process to have a clear and objective understanding of the kinds and degrees of linguistic competence that are at issue in functional language use. Although one's initial contact with the FSI-type interview scale (and the sudden awareness that the "best students" in a particular class or program may be at or near the bottom of this scale) may be rather sobering, increasing familiarity with the basic concept of externally-referenced language performance as exemplified in the FSI and similar competency descriptions will allow instructors and others concerned to begin speaking what is not only a common language among themselves with respect to instructional goals and attainments but also a language comprehended by parents, school board members, legislators, prospective employers, interested citizens, and others who are not part of the educational system per se but who have a very legitimate interest in the functional outcomes of second language instruction. Making the classroom endeavor more closely congruent with real-life language use expectations is an undertaking that can be very much advanced through broader dissemination and use of proficiency standards such as those represented by the FSI verbal descriptions and by periodic assessment of student progress along this overall competency scale.

With regard to the second element, the time is highly appropriate to dismiss from the concern of the teacher the laborious and technically difficult preparation of multiple-choice or other discrete-point tests of individual language elements, as well as to put aside, at least for classroom testing purposes, further involvement

with cloze passages, reduced redundancy tests, and other measurement approaches that do not directly reflect the bulk of real-life language use situations. Instead on the assumption that the instructional process itself will, over the foreseeable future, continue to have as a primary goal the development of functional language competency - the two primary endeavors of the classroom teacher should be; 1) present and to have students actively practice using the target language in situations that reflect as closely as possible the real-life contingencies at issue in these situations; 2) to monitor student attainment of these competencies through the use of testing procedures which themselves represent real-life language use to the greatest feasible extent. The previous cited work of Omaggio and a related just-published text by the same author provide useful guidelines in this regard but the considerations at issue are in all probability already quite familiar to teachers who are already following a "proficiency" orientation reasonably closely in their own teaching activities.⁴⁵ Additional resource materials addressing proficiency-oriented classroom testing, as well as journal articles and workshops at professional meetings, will help to explicate further this assessment approach, and ACTFL may be able to play a major role in such a dissemination effort as a natural extension of its work in the direct proficiency testing area.

With respect to the third element, close involvement of the teacher as an informed consumer of computer-mediated information on student learning performance must await the initial development of the relevant CAI teaching and testing materials. However, to the extent that teachers and other front-line persons involved in the teaching process can become familiar with the computer and its general capabilities even while such materials are being developed, the required "learning curve" will be abbreviated and the likelihood of successfully integrating computer-assisted instruction/assessment as an integral component of the total instructional system will be considerably increased.

NOTES

1. Bernard Spolsky. "Linguists and Language Testers."

Approaches to Language Testing. ed. Bernard Spolsky, (Arlington, VA: Center for Applied Linguistic. 1978).

2. See Robert Lado, *Linguistics Across Cultures* (Ann Arbor: University of Michigan Press. 1957); *Language Testing. The Construction and Use of Foreign Language*

Testers (London: Longmans, 1961)

3 See especially the Contrastive Structure Series. Edited by Charles A. Ferguson and published by the University of Chicago Press from 1962-65.

4 Robert Lado (note 2 above), p. 20.

5 John B. Carroll. "Fundamental Considerations in Testing for English Proficiency of Foreign Students" (Washington: Center for Applied Linguistics, 1961), pp. 30-40.

6 for an overview description. See Miriam M. Brian. "The MLA-Cooperative Foreign Language Tests: Tests with a New Liik and a New Purpose." DFL Bulletin, 6(December 1966), pp. 6-8. See also the review by John L. D. Clark, "MLA-Cooperative Foreign Language Tests" Journal of Educational Measurement, 2 (1965), pp. 234-44.

7 See Wilmarth H. Stair. "MLA Foreign Language Proficiency Tests for Teachers and Advanced Students." PMLA, 77 (1962), pp. 1-12.

8 Wilson L. Taylor, "Cloze Procedure: A New Tool for Measuring Readability." Journalism Quarterly, 30 (1953), pp. 414-38.

9 John W. Oller, Jr., "Scoring Methods and Difficulty Levels for Cloze Tests of Proficiency in ESL." Modern Language Journal, 56 (1972), pp. 151-58.

10 John W. Oller, Jr. & Nevin Inal. "A Cloze Test of English Prepositions." TESOL Quarterly, 5 (1975), pp. 37-49.

11 Jon Jonz. "Improving on the Basic Egg: The Multiple Choice Cloze Test." Language Learning, 26 (1976), pp. 255-65.

12 See Robert B. Kaplan & R. A. Jones, "A Cloze Procedure Test of Listening for University Level Students of ESL" (Los Angeles: Univ. of Southern California, 1970) (mimeo); Barry L. Nutter, "Presentation Methods, Deletion Patterns, and Passage Types for Use with Aural Cloze" Diss., Univ. of Arizona, 1974.

13 John W. Oller, Jr., Research with Cloze Procedure in Measuring the proficiency of non-native Speaker of English: An Annotated Bibliography (Arlington, VA: Center for Applied Linguistics, 1975).

14 The Cloze Procedure: A Selected Annotated Bibliography, comp. Pamela M. Reilley (Lae: PNG Univ. of Technology, 1973).

15 John W. Oller, Jr., "Dictation as a Device for Testing Foreign Language Proficiency." English Language Teaching, 225 (1971), pp. 254-59.

16 Bernard Spolsky et al., "Preliminary Studies in the Development of Techniques for Testing Overall Second Language Proficiency." Language Learning (Special Issue 3, August 1968), pp. 79-101.

17 For an overview and selected bibliography. see Pardee Lowe, Jr. & Judith E. Liskin-Gasparro. "Testing Speaking Proficiency: The Oral Interview" (ERIC/CLL Q & A series) (Washington: Center for Applied Linguistics, n.d.).

18 Claudia P Wilds. "The Oral Interview Test." Testing Language Proficiency. Ed. Randall L. Jones & Vernard Spolsky (Arlington, VA: Center for Applied Linguistics, 1975), pp. 29-44.

19 Howard E. Sollenberger. "Development and Current Use of the FSI Oral Interview Test." Direct Testing of Speaking Proficiency Theory and Application. Ed. John L. D. Clark (Princeton Educational Testing Service, 1978.)

20 Direct Testing of Speaking Proficiency Theory and Application. Ed. John L. D. Clark (Princeton Educational Testing Service, 1978)

21 An informative report of the nature and outcomes of these conferences is given in Howard T. Young. "On Using Foreign Service Institute Tests and Standards on Campuses" Measuring Spoken Language proficiency. ed. James R. Fruth, (Washington: Georgetown Univ. Press, 1980.)

22 Alice C. Omaggio. "Priorities for Classroom Testing for the 1980s." Proceedings at the National Competent on Professional Priorities. Ed. Dale L. Lange (Hastings-on-Hudson, NY: ACTFL, 1980), pp. 47-53.

23 Omaggio (note 22 above), p. 49.)

24 Susana Brutch. "Convergent/Discriminant Validation of Prospective Teaching Proficiency in Oral and Written Production of French by Means of MLA Foreign Language Proficiency Tests for Teachers (TOP and TWP) and Self Ratings." Diss., Univ. of Minnesota, 1979.

25 Renate A. Schulz. "Discrete Point Versus Simulated Communication Testing in Foreign Languages," Modern Language Journal, 61 (1977), pp. 94-101.

26 Solveig Olsen, "Foreign Language Departments and Computer-Assisted Instruction: A Survey." Modern Language Journal 64 (1980), pp. 341-49.

27 Richard T. Scanlan. "Computer-Assisted Instruction in Latin." Foreign Language Annals, 13 (1980), pp. 53-55.

28 Robert M. Terry, "Students Work with MONIQUE and Learn French," Foreign Language Annals, 10 (1977), pp. 191-97.

29 G. E. Nelson et al.. "Two New Strategies for Computer-Assisted Language Instruction (CALI)." Foreign Language Annals, 9 (1976), pp. 28-37.

30 Leon I. Twarog & E. Garrison waiters. "Mastery-Based, Self-Paced Instruction in Foreign Languages at Ohio State University," Modern Language Journal, 65 (1981), pp. 1-23.

31 National Center for Education Statistics.

"Instructional Use of Computers in Public Schools (Document NCES 82-245) (Washington: US Department of Education, 1982).

32 John H. Harrison. "Foreign Language Computer Software: What? Where? How Good?." Northeast Conference Newsletter, 13 (1983), pp. 26-30

33 John L. D. Clark, "Psychometric Considerations in Language Testing." Approaches to Language Testing, ed. Bernard Spolsky (Arlington, V-I: Center for Applied Linguistics, 1978), pp. 15-30.

34 See Marianne L. Adams. "Measuring Foreign Language Speaking Proficiency: A Study of Agreement Among Raters." Direct Testing of Speaking Proficiency: Theory and Application. ed John L. D. Clark (Princeton: Educational Testing Service, 1978), pp. 129-49; Karen A. Mullen. "Rater Reliability and Oral Proficiency Examinations." Proceedings of the First International Conference on Frontiers in Language Proficiency and Dominance Testing. ed. James E. Redden (Carbondale: Southern Illinois Univ., 1977).

35 Ray T. Clifford. "Reliability and Validity of Language Aspects Contributing to Oral Proficiency of Prospective learners of German." Direct Testing of Speaking Proficiency Theory and application. Ed. John

L. D. Clark (Princeton Educational Testing Service, 1978), pp. 193-209

36 Lale F. Bachman & Adrian S. Palmer. "Convergent and Discriminant Validation of Oral Language Proficiency Tests." Proceedings of the the First International Conference on Frontiers in Language Proficiency and Dominance Testing. ed. Raymond Silverstern. Carbondale: Southern Illinois Univ., Press. pp 53-62.

37 Rav T Clifford. "foreign Service Institute factor Scores and Global Ratings." Measuring Spoken Language Proficiency. Ed. James R. Fruth (Washington: Georgetown Univ., Press. 1980), pp. 27-30.

38 John L. D. Clark. "Interview Testing Research at Educational Testing Service." Direct Testing of Speaking Proficiency Theory and Application. Ed. John L. D. Clark (Princeton: Educational Testing Service, 1978), pp. 211-28.

39 Bruce Fraser, Ellen Ratell & Joel Walters. "An Approach to Conducting Research on the Acquisition of Pragmatic Competence in a Second Language." Discourse Analysis in Second Language Research. Ed. Diane Larsen-Freeman (Rowley, MA: Newbury House, 1980), pp. 75-91.

40 Michael Canale & Merrill Swaim. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing." Applied Linguistics. (1980), pp 1-47.

41 David E. Ingram & Elaine Wyhe. Australian Second Language Proficiency Rating. (ALSPR) (Vrisbane Mt. Gravatt College of advanced Education, 1981) See also David E. Ingram. "The Australian Second Language Proficiency Rations: Their Nature, Development, and Trialling." Directions in Language Testing. ed. John A. S. Read (Singapore Univ., Press. 1981), pp 108-36.

42 John B. Carroll et. al.. The Foreign Language Attainments of Language Majors in the Senior Year. A Survey Conducted in US Colleges and Universities (Cambridge: Harvard Univ., Graduate School of Education, 1967)

43 Copies of the expanded scale descriptions are available on request from the American Council on the Teaching at Foreign Language. Box 40. Hastings-on-Hudson, NY 10706

44 Randall L. Jones (personal communications).

45 Alice C. Omaggio, Proficiency-Oriented Classroom Testing (Washington: Center for Applied Linguistics, 1983).