LANGUAGE PROFICIENCY INTERVIEW TESTING: AN OVERVIEW

David J. Keitges
Nanzan University

ABSTRACT

This paper explores and discusses the usefulness and limitations of language proficiency interview testing. The prominent testing model is described and critiqued. Limitations of this model and this type of testing are discussed and the resultant modifications are surveyed. A process for the development of interview tests which are valid, reliable and practical is proposed and explained, and numerous practical suggestions concerning the designing, administering and rating of such interview tests are made.

I INTRODUCTION

"Language as communication" is a familiar theme found in journal articles, textbook prefaces, and in discussions among teachers and students of English in Japan today. By relating the two, writers and discussants are typically trying to redefine what language teaching — and language learning — must be concerned with — the acquisition of communicative competence in social settings. Learning another language as a communicative tool is hastened by a variety of factors, including academic use, commercial and technological uses in international trade, and touristic and other needs. Learning foreign languages in Japan has also become related to notions of "internationalism" and "intercultural communication." Language learning, then, has widened in scope and become a more practical endeavor than ever before.

Viewing language learning in such practical terms has focused critical attention on language education and its aspects: needs assessment, program design, methodology, evaluation. Increasingly, questions about language learning phrased by students, teachers, and others, relate to the authenticity and practicality of the language taught and the efficiency of the methods employed. Students want to know how they can learn "real" English, and teachers are attempting to bring "the real English-speaking world" into the classroom itself. Research has become preoccupied with the tasks of defining and describing needs, mounting courses, and evaluating progress.

Further, viewing language as a communicative tool requires increased sophistication in this evaluation. Assessing only a student's ability to manipulate discrete, separate elements of language on paper gives the teacher and administrator information of dubious value. A practical ability in speaking a second/foreign language demands evaluative procedures that can certify progress, diagnose weakness, and predict performance. Proficiency testing, therefore, is an essential part of any course or program which claims to teach language for use.

This paper will explore and discuss the usefulness and limitations of language proficiency interview testing, or, more specifically, the direct testing of speaking proficiency in the interview setting. Readers in mind are those of the English language teaching community, both native and non-native speakers, who teach language for communicative use. Also included are those who assess speaking proficiency on a more formal basis in company-wide, governmental and other programs in which a higher proficiency rating can lead to promotion, an overseas post, a certificate for employment and other career-related opportunities. After definitions of terms and anecdotes of recent experience, a model of language proficiency interviewing will be described and critiqued in Section II. Its limitations will be discussed in Section III and modifications of the model will be surveyed. Section IV will provide guidelines and suggestions for designing, administering, rating, and improving language proficiency interview tests.

Clark (1979, p. 36) distinguishes between direct, semi-direct, and indirect measures of speaking ability. Direct measures include all those procedures in which the "examinee is asked to engage in a face-to-face communicative exchange with one or more human interlocutors." Indirect measures require no active speech production by the examinee but depend on paper-and-pencil varieties of cloze tests and other such "productive" techniques. Semi-direct measures of speaking ability elicit active speech but by means of tape recordings, printed test booklets and other "nonhuman" elicitation procedures. Only direct measures involve actual oral exchanges and contain what Carroll (1980, p. 54) calls oral interaction: "constructive interplay with unpredictable stimuli."

Further, for our purposes, the evaluation of an individual's language proficiency must be distinguished from that same individual's language achievement. Proficiency refers to overall or global competence in a language, regardless of how that competence was acquired. Achievement evaluation, on the other hand, measures an individual's acquisition of specific linguistic features of the language that have been presented in, for example, a particular language course. (Clark 179, p. 39) Therefore, we can define such language proficiency interviews as described above as evaluative sessions in which one or more persons perform communicative tasks elicited or assigned by one or more examiners who subsequently observe and rate the resultant speaking performance to determine speaking proficiency.

Anecdotal comments relayed from examining partner to partner and from colleague to colleague frequently pose rather significant questions about the usefulness and limitations of language proficiency interview testing. Some examiners complain that judging speaking proficiency is so subjective that the results of such tests are all too suspect; others reject such interview testing on the grounds that the "real" elements of language (grammar, vocabulary, etc.) cannot be fairly tested in this manner. A second category of difficulty related to interview procedure. Examiners sometimes aren't sure how they are to elicit speech samples

and exactly what they are to evaluate: the whole, the parts, the overall impression, what. And examiners working together who do not discuss criteria of evaluation beforehand often rate the proficiency of the same individual in very different ways. And even when checklist scales are used during interviews, some examiners question if "dismembering the discourse" is an appropriate way to judge the communicative proficiency of the interviewee. These concerns and numerous variations are frequently expressed. How, then, are we to regard the element of subjectivity in language proficiency interview testing? And what procedures can aid us in eliciting appropriate speech samples and evaluating them efficiently? These and other questions will be discussed in Sections II, III and IV below.

II A TESTING MODEL: DESCRIPTION AND CRITIQUE

In the minds of many, the direct testing of speaking proficiency by interview is nearly synonymous with the procedure developed by the U.S. Foreign Service Institute (FSI). Most teachers and testers have at least a passing familiarity with this procedure as it has been widely though incompletely described in major teaching methods and testing texts (Valette 1977, p. 157-51; Rivers 1981, p. 368, 497-99). As much of the research and many of the modifications discussed in the literature are related to this particular testing model, perhaps it is wise to explore it briefly to distinguish its specific purpose, target population, and procedures.

Wilds (1975, p. 29-44) has provided the most comprehensive description of the FSI interview test. The procedure, first adopted in 1956 and somewhat revised since then, was devised to assess the foreign language speaking abilities of U.S. Government personnel, especially diplomatic, military and aid officials, and, later, Peace Corps volunteers. The evaluation takes place in an interview conducted by two examiners for a sole interviewee. The trained examiners engage the interviewee in a conversation to determine his speaking proficiency level as defined by the functionally-based FSI Proficiency Ratings (Appendix A). In FSI usage,

the team of examiners consists of a senior member, a native-speaking certified language examiner or a linguist thoroughly familiar with the target language, and a native-speaking junior member who elicits the samples of speech to be evaluated. The senior member, though an interested participant, usually refrains from actual elicitation. The speech samples of the interviewee are judged to range from elementary proficiency through limited working proficiency, minimum professional proficiency, and full professional proficiency, to native or bilingual proficiency. These five proficiency levels, which are further divided with plus (+) notations to indicate half-level proficiencies, were devised in response to the specific needs of the testing population.

The taped interview begins with simple social formulae which, if handled badly, leads the examiner to put a ceiling on the difficulty of subsequent questions (and thus the rating level) of the interviewee. If the initial queries are dealt with satisfactorily, the examiner will go on to more difficult topics of an autobiographical or professional nature. Informal oral interpretation between examiners is sometimes assigned to elicit certain desired grammatical or lexical items unused by the interviewee. The interview ranges from fifteen to less than thirty minutes in length. Final rating occurs either during the interview or directly after its conclusion. The examiners may (but are not required to) use a "Checklist of Performance Factors" (accent, grammar, vocabulary, fluency, comprehension) and a weighted conversion table (Appendix B) to reach their conclusions. The examiners may also report specific weaknesses to the interviewee at the interview's end or fill in a "Factors in Speaking Ability" chart designed for the same purpose (Clark 1979, p. 40).

In terms of validity, reliability and practicality, the three imperatives of test design and administration, the FSI oral interview appears to be an excellent measure of the speaking proficiency of the target population. Consisting of an interview in which oral interaction occurs, it holds both *face* and reasonable *content* validity as it "measures what it is supposed to measure"—the ability of the person to speak the language (Clark 1979, p. 37). It has also demonstrated high

statistical reliability (Oller 1979, p. 392; Hendricks, et al. 1980, p. 78). The interview is also practical, at least for the FSI, as it can be administered, rated and interpreted with ease (Wilds 1975, p. 29-30).

There are a number of practical factors, however, which underlie the validity and reliability of this procedure. Oller (1979, p. 326) states that interview validity (and hence reliability) depends on three factors: (1) how the speech acts are elicited; (2) what the rating scale(s) are referenced against (the criteria chosen for deciding proficiency); and (3) who is a qualified rater. It is in these practical aspects, the details of the procedure, that the reasons for its success can best be viewed.

The *Manual for Peace Corps Language Testers* (1970, p. 11) defines the purpose of the language proficiency interview in this way: "It is *not* simply a friendly conversation on whatever topics come to mind . . . It is rather a specialized procedure which efficiently uses the relatively brief testing period to explore many different aspects of the student's language competence in order to place him into one of the categories described." This definition emphasizes the importance of the elicitation of speech samples, the grading of them in reference to the predetermined proficiency criteria and ratings, and, further, implies the necessary training of the human administrators who must elicit and rate skillfully and reliably. These procedural requirements are fulfilled in various ways at the FSI. Wilds (1975, p. 34) reports that, for languages that are tested often (more than 60 languages are tested), there are libraries of tapes of previous tests of all levels for the training rater to use. In addition, there is a "substantial amount of written material aimed at clarifying standards and suggesting appropriate techniques." A much experienced staff also helps guide others in the achievement and maintenance of testing competence. But, it may be asked, does this training and the accumulated experience of the examiners, though impressive, fully explain the success of the FSI interview? What, if anything, underlies these aspects of interview test success?

The element of subjectivity, that of the examiner's per-

sonal judgement, present in judging oral performances has long been pinpointed by some as a major disadvantage, and one that inhibits its more extensive use. Without doubt, the examiner's personal judgement is present in the eliciting and rating of speech samples. It is obvious that the examiner must inject himself into the test by deciding which questions to ask and how to phrase them, and how the multitude of possible responses are to be rated in accordance with the still loosely-defined criteria. But how shall this subjectivity be regarded? As a disadvantage, even an obstacle? Or as a benign or even beneficial component of the process?

Much has been done at the FSI to limit the influence of subjective judgement on testing: the Proficiency Ratings, based on functional use, have been devised; examiners have learned techniques of elicitation that furnish speech samples considered valid and appropriate for rating; raters are carefully trained to differentiate good, fair, and poor performances reliably according to defined criteria. Subjectivity, then, or at least its more negative influences, has been limited. Rivers (1981, p. 69) observes that oral evaluation is essentially subjective but notes that raters can be trained—as they are at the FSI—to reach basically comparable results. Oller (1979, p. 328) defines oral evaluation as judging "subjectively according to loosely stated criteria." Further, Oller (p. 48) notes that objective procedures in evaluating oral performance are not necessarily more reliable than subjective ones. "Certain aspects of language performances may simply lend themselves more to subjective judgement than they do to quantification by formula." Clark (1979, p. 41) agrees by noting studies that demonstrate that subjective rating of oral performances should not be regarded as "intrinsically unreliable." In addition, it may be unreasonable and unwise to ask that rating of oral performances be accomplished in a more "objective" manner. For, considering the paucity of our present knowledge and understanding of oral interaction, how and in what form would objective rating take place? In fact, instead of asking if the element of subjectivity is appropriate or not in proficiency rating, it may be more reasonable and useful to ask to what extent subjectivity,

granted an informed varient, is necessary to the process. For this reason, it may be concluded that the FSI interview test is successful not only because of its established criteria and rating system and the careful training of its examiners, but also precisely because of the controlled or informed subjectivity exercised by those examiners during the interview process.

III LIMITATIONS AND MODIFICATIONS

Up to this point, the design, procedure and rating system of the FSI interview test has been briefly reviewed. Its usefulness in assessing speaking proficiency has been outlined. Also, the reasons for its success have been explored. From now, attention will be shifted to the limitations of language proficiency interviewing, and the FSI model in particular. Also, modifications of the model will be surveyed to demonstrate how teachers and researchers have dealt with these recognized limits of use.

Language proficiency interviewing leaves much to be desired. The described FSI procedure, although efficient with the specific group for which it was designed, has a number of limitations in its general applicability. Chief among these is its absolute scaling of performances. Absolute scaling requires reliable, consistent judgements by different examiners over time. Individuals tested today must have their performances rated under the same conditions as those rated previously, and those who will be rated in the future. The essence of absolute scaling, then, is consistency. Relative or normative scaling, on the other hand, seeks only to differentiate among members of a particular group at a particular time. There is no specific, absolute need for consistency over time. Relative scaling, in the classroom and at the employment agency, is the normal method, while absolute scaling is only for rare, special situations. The difficulty of the matter becomes apparent when one considers the training of the examiners charged with the rating responsibility. The need to maintain reliability over time further burdens, and strictly formalizes, the training of FSI examiners. Other, relative

raters need not be so burdened. In addition, while the increased amount of this examiner-training and lengthy interview sessions may not be prohibitive for the FSI and other select groups, these requirements are burdensome for other testers. And, though the basic 5-level Proficiency Ratings discriminate abilities well enough for the FSI, these same ratings do not distinguish abilities well enough for secondary school and college learners with more limited proficiency (Reschke 1978, p. 80–1). Finally, the inclusion of listening comprehension as a checklist performance factor by the FSI is seen by many as undesirable as examiners, no matter how skillful, are exposed to so little evidence of comprehension that it is unlikely a fair and complete assessment of this important skill can be made in an interview alone. Therefore, although the FSI model is suitable for its designed purpose, it cannot be freely transferred to dissimilar testing situations. The implicit (and sometimes explicit) recommendations for its unmodified use by some writers need to be softened and qualified.

Other sorts of limitations of the FSI model have been indicated by "integrative" and "pragmatic" testers. This approach, backed by Carroll (1980), Cohen (1980), and Oller (1979), among others, differs most significantly from the FSI model in the scoring and rating of speech samples. Oller (p. 305) believes that scoring techniques used in interviews should relate not only to morphology and syntax as now, but also to the overall meanings contained in utterances. Thus, he calls for a wider array of criteria for judging speech beyond the traditional FSI set (accent, grammar, vocabulary, etc.). Other researchers (Callaway 1980, p. 111; Hendricks et al. 1980, p. 85; Mullen 1980, p. 101) have statistically examined the present FSI scoring techniques and found them wanting. This research suggests that using integrated or unitary scaling for scoring interviews, rather than the FSI multiple checklist scales, is preferable because there is not adequate evidence to prove that the FSI checklist scales actually measure different things. Mullen, in particular, noted that an overall scale of proficiency appeared to represent a composite of the four other scales in her research and was, statistically, more

reliable—thus preferable. Qualified findings on unitary/multiple scaling has been reported by Bachman and Palmer (1981, p. 67). Callaway, in researching raters, found that the overall comprehensibility of speech behavior is what motivates examiner's evaluations of proficiency, thus further supporting arguments for unitary scaling in interview scoring techniques. In addition, these "pragmatic" critics also encourage the use of other criteria (naturalness, clarity, suitability) in the rating of speech (Cohen, 1980, p. 20–23). Rating an individual's performance by such criteria as these, they argue, is pedagogically more useful and communicatively more accurate. An extended 9-level interview assessment scale of this type, designed by Carroll (1980, p. 135) is included as Appendix C.

Finally, in a general way, interview testing has important limitations that every tester must be aware of. Although interviews are planned to replicate as closely as possible everyday communication, in several ways they are atypical. Clark (1979, p. 38) has observed that "talking to the examiner isn't the same as speaking to a waiter, taxi driver or friend." The psychological and affectional components of communication present in the interview also differ from ordinary communication (Jones 1975, p. 14). In addition, the usefulness of interview-based evaluations may be lessened by differences in communication style transferred from another language and culture by the interviewee to the target language. Richards (1981, p. 7–26) details some of the problems that occur in conversations as a result of this communication style transference. These recognized limitations, however, do not invalidate interview tests as measures of speaking proficiency. They merely require that such interviews be designed and conducted with the utmost care and flexibility necessary.

Modification of a model is a natural consequence of recognized inadequacy. And, as the FSI model is limited in a number of practical ways, changes in its design and use have readily been made. These changes, briefly surveyed, come in six areas. The first, the use of a sole examiner instead of two at the FSI, is unavoidable for most teachers and testers due to the number of interviewees and the burden

of other work. Further, examiner self-training (through reading, experience, heightened awareness) is a related accommodation to practical circumstances. The third common change, a shortening of interview-time, is also an obvious accommodation. Though the FSI can extend interviews to nearly half an hour, few teachers or testers can possibly do so. This has inevitably resulted in interviews being conducted in between five and ten minutes per interviewee. It is important to note here that the Educational Testing Service (Clark 1978, p. 227–8) has determined that interviews in the five-to-seven minute range are adequate for valid and reliable rating by trained examiners. The fourth common change is to increase the efficiency of the interview by testing groups instead of individuals. Reschke (1978, p. 82) has suggested testing from three to five persons at each session. The present writer has much successful experience with groups of three. The fifth general modification concerns the checklist and rating scales. Clark (in Valette 1977, p. 161) has simplified and shortened the FSI checklist scales to make them more useful in the classroom. Schulz and Bartz have developed "pragmatic" scales that score the level of communication present in the interviewee's speech sample. The Schulz Communicative Competence Scale (in Valette 1977, p. 161) rates "fluency," "comprehensibility," "amount of communication," and "quality of communication." Bartz' scale (in Valette 1977, p. 150–1) mirrors Schulz' except that "effort to communicate" is interchanged with "comprehensibility" (Appendix D). The sixth and final major area of change from the FSI model is the specific use of the interview for other purposes. Although the majority of these other uses will be outlined in Section IV, the use of the interview for strictly diagnostic purposes has been discussed by Graham (1978, p. 33–9).

These modifications demonstrate that the general format of the FSI oral interview is adequate as a model for the design of language proficiency interviews. Though the FSI model is strictly appropriate for only the specific purpose for which it was designed, it can be modified to serve wider

testing requirements. Properly understood, the FSI interview model can aid both classroom teachers and more formal testers in their design, administration, and rating of proficiency interview tests.

IV A PRACTICAL PRIMER: GUIDELINES AND SUGGESTIONS

The purpose of this section, plainly, is to provide some guidelines and practical suggestions for the designing, administering, and rating of language proficiency interviews. The three former sections presented a rationale for this kind of testing, described and critiqued the most prominent model, and surveyed limitations and modifications. This section is to be a practical conclusion. But, before launching into practicality, I would like to extend the discussion to include the many practical uses of proficiency interviews. These popular uses went largely unmentioned before in order to avoid unnecessary confusion during the description and critique of the FSI model. Now it aids in the effort to recognize these other purposes: the language proficiency interview (LPI) as a part of a formal or informal course or program needs assessment study; the use of the LPI in ability-grouping and as a motivational and self- and peer-grading technique; the LPI as a diagnostic evaluation instrument and as a practical means of establishing and maintaining classroom and program goals of communicative language use; and the use of the DPI to maintain certain course or program proficiency requirements or standards (foreign-language teacher certification, etc.). These common uses demonstrate that interviewing can also be instrumental means to various practical teaching and administrative ends.

The design of an appropriate LPI procedure implies more than the mere adoption of a recommended rating scale. In fact, proper design requires the employment of a process than can ensure a valid, reliable and practical test instrument. The FSI model, itself, is the successful product of just such a design process—a process others can use to create proficiency interview tests suited to their particular circumstances.

This process of design is the main subject of discussion below.

Ryan and Frederiksen (1951), cited and discussed by Jones (1979, p. 52-3), developed a process for the preparation of performance tests which are valid, reliable and practical. Though intended as a general process for a wide variety of testing needs, it can be adapted to the more specific needs of language proficiency interviewing. Ryan and Frederiksen listed seven steps in performance test design: (1) make a job analysis; (2) select tasks to represent the job; (3) develop a rating form; (4) survey the practical limitations; (5) develop a tentative operating plan; (6) try out the test and revise it; and (7) prepare directions for administration and use of the test.

This general process can be adapted to serve LPI testing. By re-wording and grouping some of the steps, the process can be reduced to five easily remembered stages: (1) analyze the needs; (2) select representative tasks; (3) develop a rating form; (4) accommodate the limitations; and (5) train the examiner. By following these steps carefully, the test designer can devise an instrument for the specific group of learners which takes into account all of the variables of their language needs and use, as the FSI instrument does for those of its target population. These stages of test development are discussed separately below.

1. *Analyze the communicative needs of the interviewee(s)*

For what reason is the interviewee being tested? To decide if his speaking proficiency is adequate or not for a certain job? To rank his ability in relation to his fellow learners? Questions of this type are critically important at this first stage. Before it is possible to decide *how* to conduct the interview, the need for which the interview is being held must be specified. Needs range from the communicatively narrow (taking customers' orders in a restaurant) to the very broad and complex (negotiating contracts for the purchase of computer components). Needs can also be teacher-derived (diagnosis) and program-centred (certification requirements). Once this need is specified clearly, the process of designing the interview test can proceed smoothly; without its description,

the testing will be a hard and likely fruitless task.

Practically speaking, this need should be written down as the purpose statement of the test, for the designer's later reference when he is planning how to elicit and rate the speech samples. This written description should also include information relating to the target communicative settings (formal/informal business meetings, academic seminars, touristic exchanges, etc.), needs for expert knowledge (scientific, commercial, etc.), other special requirements (speech-making, for instance), and the tolerance of error allowable in the interactions. Further, the statement should list at least some of the functional uses of language that will be required. Most teachers and testers are familiar with the writings of D. A. Wilkins (1976) and J. A. van Ek (1976) on the notional and functional uses of language in communication. Van Ek (p. 45–49) proposes a list of functions (accepting an offer or invitation, expressing capability or incapability, expressing disappointment, etc.) that can aid in this need specification of the interviewee. John Munby (1978, p. 123–131) also provides a taxonomy of language skills which can be consulted for this purpose.

2. *Select representative communication tasks for testing*

At this stage, a representative sample of the communication tasks of the targeted need must be selected. Obviously, the interviewee cannot be tested for all of the needs that have been identified. Therefore, a small but appropriate sample must be chosen for the brief testing period. Sampling should range from the easy to the more difficult functional tasks (greeting, apologizing, explaining, defending, etc.) and deal with the specific areas and topics of discourse that have been pinpointed. Therefore, when interviewing electrical engineers, for example, some of the tasks must relate to true-to-life engineering topics, possibly the individual's special field or research. Asking only social/general questions of an individual being tested for engineering-specific speaking proficiency results in inadequate knowledge of his true proficiency and a wasted interview.

The second challenge of interview test design, then, is

deciding what questions to ask and how to ask them. In a conversational format, two strategies for efficient questioning can be recommended. The first is the use of carefully prepared "ceiling questions" which are designed to test the interviewee's proficiency at the top of each level of the proficiency rating scale. Appropriate greetings and such opening formalities may convince the examiner that the interviewee possesses sufficient proficiency to be tested on the next higher level. Responses to subsequent "ceiling questions" may suggest that a certain level should be further explored for difficulty. When this tentative rating level has been found, the second strategy, that of asking several related questions of increasing difficulty on a specific topic, may be used. The purpose of this strategy is to make certain the rating level and explore the breadth of the interviewee's vocabulary and his general ability to engage in topical discussion. It has been noted by Morrow (in Carroll 1980, p. 12) that language is essentially interactive, unpredictable, purposive and contextualized. Therefore, the greatest care should be taken to ensure that the communicative exchanges initiated by the examiner replicate as closely as possible genuine language use. Both types of questions, in sufficient quantities, should be prepared ahead of time and written down for instant reference.

Direct interview testing, as defined in Section I, also includes the use of role-play, informal oral interpretation, reversal of roles, situational problems, group discussions on prepared topics, and other techniques. These techniques are especially useful when more than ten minutes is available for each interview, and when group interviews are held. I have found that "interviewing" groups of three students engaged in a pre-planned discussion without notes for thirty minutes is both a useful and enjoyable way to assess speaking proficiency.

Speeches, both extemporaneous and prepared, however, should not be assigned as proficiency evaluation tasks. This communication form, though oral, lacks the essentially inter-

active nature of genuine oral communication.

As long as the questions asked and the tasks assigned are directly related to the needs of the interviewees and are not so contrived as to tax the imagination unduly, these approaches to eliciting speech samples will be successful.

3. *Develop a Rating Scale*

Much has already been written about checklist and rating scales, and examples have been placed in the appendix for the reader's review. Jones notes in his discussion (1979, p. 53) that "the key to achieving objectivity in a performance test is the checklist and rating scale." While this may be true of non-language tests, we have seen that clear objectivity in interview tests cannot be achieved completely due to the variety and complexity of oral communication. Nonetheless, predetermined rating criteria must be used to "channel" the examiner's attention toward those factors considered most important in this communication. Criteria chosen for general consideration should include not only the traditional set (accent, grammar, vocabulary, fluency, and, to the degree possible, comprehension) but also the so-called criteria of communication: quality and amount of communication, effort to communicate, and communicative effectiveness. In my experience, it is best to place emphasis on the latter group and allow the former set to qualify or explain the conclusions reached. Thus, I value the extent to which an individual can communicate his desires clearly in the language more than his excellence in pronunciation. Pronunciation, in fact, should only be considered a factor when it detracts from the interviewee's communicative ability in a specific setting. Vocabulary and grammar, likewise, become significant when the specified topics raised cannot be discussed with accuracy and in sufficient detail.

In attempting to judge communicative effectiveness, it is useful to qualify one's overall impression with the interviewee's use of rhetorical behaviors. These behaviors come into play as the proficiency of the individual increases. Einhorn (1981, p. 217-228) has isolated six such behaviors

(identification with interviewer, argument support, organization, style, delivery, and images conveyed) that determined success or failure for applicants in job interviews. Such behaviors are also significant factors in the effectiveness of interpersonal oral communication.

How one uses these chosen (and weighted) criteria in arriving at rating decisions is much in dispute. While the formal FSI model suggests that final rating is decided by the computation of five separately scored factors, in fact FSI examiners rarely use this procedure. Apparently, they are so familiar with the characteristics that differentiate one rating level from another that they can easily categorize performances. Whether they even analyze performance in terms of the five factors specifically is uncertain. Callaway (1980, p. 111) states that dividing oral performance into components is superfluous as raters make holistic unidimensional judgements. My experience suggests that it is unnecessary to compute separate scores but that intimate knowledge of the rating levels is necessary so that the rater can transfer his essentially subjective judgements to that scale efficiently. Criteria, written down as reference, are essential to maintain the examiner's approach to performance rating for all interviewees. Practical knowledge of the rating scale grounds the criteria in function.

With the foregoing in mind, a rating scale must be developed that relates directly to the interviewees. The FSI scale, gross in units (only five) and stretching from intermediate speaking ability to more advanced, may be appropriate for few testing situations. More detailed is the business-oriented Daiei scale (Appendix E) and Carroll's scale. The latter seems good for general academic use, and language school courses. In any case, as proficiency rating charts are extremely difficult to construct, especially when several levels of proficiency must be differentiated, it is recommended that already available scales be adapted for specific purposes. With adequate experience and special need, custom-made proficiency ratings can be devised.

4. *Accommodate the Limitations*

Accommodating the limitations is the sub-process of fitting the interview test to the circumstances. At this stage of design and preparation, the test has already been shaped in a variety of ways, perhaps subconsciously, by the limits of time, ability level, urgency or importance of purpose, etc. The designer has likely already decided the length of each interview, whether individuals or groups will be tested, the amount of preparation required of the examiner, and other factors. It is useful, nonetheless, to review the form of the interview and its operating schedule to see if any improvements can be made or if anything has been forgotten. One factor that seems to be often neglected is examiner-fatigue. This is a malady which invariably attacks examiners about half-way through planned interview schedules. To lessen its effects, a regular rest break should be taken during each hour of interviewing.

5. *The Training of the Examiner*

What remains to be accomplished is the training of the examiner and the pilot-testing of the planned interview test —two requirements that complement one another rather well. Just as the interviewees must have prepared themselves over their learning period to perform successfully, so must the individual examiner prepare himself to elicit and rate well. This preparation takes two forms, general and specific. General preparation involves the examiner's intellectual understanding of the interview as a forum of communication and as a testing vehicle. This form of preparation requires study in the disciplines of testing, discourse analysis, and related areas. It also requires an increased sensitivity toward language use and a keen ability to differentiate genuine and effective oral performances. Specific training is best accomplished in the interview process itself. As noted before, the FSI uses tapes of previous tests and anecdotal materials prepared by experienced examiners in the training of new raters. Few, if any, ordinary testers have an opportunity for such a careful training. Novice testers, however, can ask to join in interviews conducted by colleagues, or do "mock"

interviews during class time to gain experience. Further, the novice can "test" the English of native-speakers and fluent non-native speakers as means of setting appropriate expectations of performance for later interviews. There are many daily opportunities to evaluate speech, and these can be used for training purposes. In fact, the design of the interview test includes "training" in needs assessment, task selection, criteria choice, and measurement scale design. In the final stage of examiner training, the planned interview test can be pilot-tested with an unrelated group of individuals to test its actual practicality and the ability of the examiner to elicit and rate efficiently. These pilot-testings can be cassette-recorded and reviewed later. Thus, in any number of ways, examiners can prepare themselves for the difficult tasks involved in interview testing of speaking proficiency.

CONCLUSION

In Japan, perhaps more than elsewhere, testing defines learning. Learners readily identify the nature and worth of the experience by the form and difficulty of the test. Therefore, in language classes where the guiding aim has been shifted to communicative use, proficiency interview testing has a special role to play.

This proficiency interviewing has much to recommend it. In the classroom, it dramatically demonstrates the goal of the activities. Learners soon realize, at the time of the diagnostic group-interview, that it is their speaking proficiency, not their memorization skills, that is important. and "framing" a course with a beginning diagnostic interview and an ending proficiency evaluation is a powerful encouragement for taking all of the classroom communication activities seriously. Oral performances, characteristically both open and personal, force learners to monitor their own progress more carefully.

Valid and reliable proficiency interviewing is within the grasp of all teachers and testers. The process of test design and examiner training, though time-consuming initially, opens up new ways of thinking about language use—and language teaching and learning.

REFERENCES

Bachman, L. R. and A. S. Palmer, 1981. "The Construct Validation of the FSI Oral interview." *Language Learning*. 31, 1, 67-86.

Callaway, Donn R. 1980. "Accent and the Evaluation of ESL Oral Proficiency." In *Research in Language Testing*. J. W. Oller, Jr. and K. Perkins. Rowley, Mass.: Newbury House Publishers, Inc., 102-115.

Carroll, Brendan C. 1980. *Testing Communicative Performance: An Interim Study*. London: Pergamon Press.

Clark, John L. D. 1978. "Interview Test Research at Educational Testing and Service." In *Direct Testing of Speaking Proficiency: Theory and Application*. John. L. D. Clark (ed.). Princeton, N. J.: Education and Testing Service, 211-228.

— 1979. "Direct vs. Semi-Direct Tests of Speaking Ability." In *Concepts in Language Testing: Some Recent Studies*. E. J. Brière and F. B. Hinofotis (eds.). Washington, D.C.: Teachers of English to Speakers of Other Languages, 35-49.

Cohen, Andrew D. 1980. *Testing Language Ability in the Classroom*. Rowley, Mass.: Newbury House Publishers, Inc.

Daiei's English Proficiency Level Breakdown. 1981. The Daiei Language Development Program.

Einhorn, Lois J. 1981. "An Inner View of the Job Interview: An Investigation of Successful Communicative Behaviors." *Communication Education*. 30, 217-228.

Graham, Stephen L. 1978. "Using the FSI Interview as a Diagnostic Evaluation Instrument." In *Direct Testing of Speaking Proficiency: Theory and Application*, 31-40.

Hendricks, D. et al. 1980. "Oral Proficiency Testing in an Intensive English Program." In *Research in Language Testing*, 77-90.

Jones, Randall L. 1979. "Performance Testing of Second Language Proficiency." In *Concepts in Language Testing: Some Recent Studies*, 50-57.

Lado, Robert. 1978. "Scope and Limitations of Interview-Based Language Testing: are We Asking Too Much of the Interview?" In *Direct Testing of Speaking Proficiency: Theory and Application*, 113-128.

Manual for Peace Corps Language Testers. 1970. Princeton, N.J.: Educational Testing Service, 11. In *Language Tests at School: A Pragmatic Approach*. J. W. Oller, Jr. London: Longman Group Limited, 324.

Mullen, Karen A. 1980. "Rater Reliability and Oral Proficiency Evaluations." In *Research in Language Testing*, 91-101.

Munby, John. 1978. *Communicative Syllabus Design.* Cambridge: Cambridge University Press.

Oller, John W. Jr. 1979. *Language Tests at School: A Pragmatic Approach.* London: Longman Group Limited.

Reschke, Claus. 1978. "Adaptation of the FSI Interview Scale for Secondary Schools and Colleges." In *Direct Testing of Speaking Proficiency: Theory and Application*. 3, 7-26.

Rivers, Wilga M. 1981. *Teaching Foreign Language Skills.* 2nd edn. Chicago: The University of Chicago Press, 184-258.

Ryan, D. G. and N. Frederiksen. 1951. "Performance Tests of Educational Achievement." In *Educational Measurement.* E. F. Lindquist (ed.). Washington, D.C.: American Council on Education, 483-92.

Sollenberger, Howard E. "Development and Current Use of the FSI Oral Interview Test." In *Direct Testing of Speaking Proficiency:* Theory and Application.

Valette, Rebecca M. 1977. *Modern Language Testing.* 2nd edn. New York: Harcourt Brace Jovanovich, Inc., 119-163.

Van Ed, J. A. 1976. "The Oral Interview Test." In *Testing Language Proficiency*. R. L. Jones and B. Spolsky (eds.). Arlington, Va.: Center for Applied Linguistics. 29-44.

Wilkins, D. A. 1976. *Notional Syllabuses.* Oxford: Oxford University Press.

APPENDIX A

The FSI Proficiency Ratings

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

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

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

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

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

APPENDIX B

The FSI Checklist of Performance Factors and Descriptions

Accent

1. Pronunciation frequently unintelligible.
2. Frequent gross errors and a very heavy accent make understanding difficult, require frequent repetition.
3. "Foreign accent" requires concentrated listening and mispronunciations lead to occasional misunderstanding and apparent errors in grammar or vocabulary.
4. Marked "foreign accent" and occasional mispronunciations that do not interfere with understanding.
5. No conspicuous mispronunciations, but would not be taken for a native speaker.
6. Native pronunciation, with no trace of "foreign accent."

Grammar

1. Grammar almost entirely inaccurate except in stock phases.
2. Constant errors showing control of very few major patterns and frequently preventing communication.
3. Frequent errors showing some major patterns uncontrolled and causing occasional irritation and misunderstanding.
4. Occasional errors showing imperfect control of some patterns but no weakness that causes misunderstanding.
5. Few errors, with no patterns of failure.
6. No more than two errors during the interview.

Vocabulary

1. Vocabulary inadequate for even the simplest conversation.
2. Vocabulary limited to basic personal and survival areas (time, food, transportation, family, etc.).
3. Choice of words sometimes inaccurate, limitations of vocabulary prevent discussion of some common professional and social topics.
4. Professional vocabulary adequate to discuss special interests; general vocabulary permits discussion of any nontechnical subject with some circumlocutions.

5. Professional vocabulary broad and precise; general vocabulary adequate to cope with complex practical problems and varied social situations.
6. Vocabulary apparently as accurate and extensive as that of an educated native speaker.

Fluency

1. Speech is so halting and fragmentary that conversation is virtually impossible.
2. Speech is very slow and uneven except for short or routine sentences.
3. Speech is frequently hesitant and jerky; sentences may be left uncompleted.
4. Speech is occasionally hesitant, with some unevenness caused by rephrasing and groping for words.
5. Speech is effortless and smooth, but perceptibly non-native in speed and evenness.
6. Speech on all professional and general topics as effortless and smooth as a native speaker's.

Comprehension

1. Understands too little for the simplest type of conversation.
2. Understands only slow, very simple speech on common social and touristic topics; requires constant repetition and rephrasing.
3. Understands careful, somewhat simplified speech directed to him or her, with considerable repetition and rephrasing.
4. Understands quite well normal educated speech directed to him or her, but requires occasional repetition or rephrasing.
5. Understands everything in normal educated conversation except for very colloquial or low-frequency items or exceptionally rapid or slurred speech.
6. Understands everything in both formal and colloquial speech to be expected of an educated native speaker.

The FSI Weighting and Conversion Tables

FSI Weighting Table

Proficiency Description ⟶	1	2	3	4	5	6	
Accent	0	1	2	2	3	4	_____
Grammar	6	12	18	24	30	36	_____
Vocabulary	4	8	12	16	20	24	_____
Fluency	2	4	6	8	10	12	_____
Comprehension	4	8	12	15	19	23	_____
						Total:	

FSI Conversion Table

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

APPENDIX C

Carroll's Interview Assessment Scale

Band	
9	Expert speaker. Speaks with authority on a variety of topics. Can initiate, expand and develop a theme.
8	Very good non-native speaker. Maintains effectively his own part of a discussion. Initiates, maintains and elaborates as necessary. Reveals humour where needed and responds to attitudinal tones.
7	Good speaker. Presents case clearly and logically and can develop the dialogue coherently and constructively. Rather less flexible and fluent than Band 8 performer but can respond to main changes of tone or topic. Some hesitation and repetition due to a measure of language restriction but interacts effectively.
6	Competent speaker. Is able to maintain theme of dialogue, to follow topic switches and to use and appreciate main attitude markers. Stumbles and hesitates at times but is reasonably fluent otherwise. Some errors and inappropriate language but these will not impede exchange of views. Shows some independence in discussion with ability to initiate.
5	Modest speaker. Although gist of dialogue is relevant and can be basically understood, there are noticeable deficiencies in mastery of language patterns and style. Needs to ask for repetition or clarification and similarly to be asked for them. Lacks flexibility and initiative. The interviewer often has to speak rather deliberately. Copes but not with great style or interest.
4	Marginal speaker. Can maintain dialogue but in a rather passive manner, rarely taking initiative or guiding the discussion. Has difficulty in following English at normal speed; lacks fluency and probably accuracy in speaking. The dialogue is therefore neither easy nor flowing. Nevertheless, gives the impression that he is in touch with the gist of the dialogue even if not wholly master of it. Marked L1 accent.
3	Extremely limited speaker. Dialogue is a drawn-out affair punctuated with hesitations and misunderstandings. Only catches part of normal speech and unable to produce continuous and accurate discourse. Basic merit is just hanging on to discussion gist, without making major contribution to it.
2	Intermittent speaker. No working facility; occasional, sporadic communication.
1/0	Non-speaker. Not able to understand and/or speak.

APPENDIX D

Bartz' Rating Scale

A.	Fluency	1	2	3	4	5	6
B.	Quality of Communication	1	2	3	4	5	6
C.	Amount of Communication	1	2	3	4	5	6
D.	Effort to Communicate	1	2	3	4	5	6

The levels of the scales are defined as follows:

A. *Fluency* (similar to the Foreign Service Institute scale)

B. *Quality of Communication*
   1. Speech consists *mostly* of inappropriate isolated words and/or incomplete sentences with just a *few* very short complete sentences.
   2. Speech consists of *many* inappropriate isolated words and/or incomplete sentences with *some* very short complete sentences.
   3. Speech consists of *some* inappropriate isolated words and/or incomplete sentences with *many* very short complete sentences.
   4. Speech consists of *hardly any* isolated words and/or incomplete sentences with *mostly* complete sentences.
   5. Speech consists of isolated words only if appropriate and *almost always* complete sentences.
   6. Speech consists of isolated words only if appropriate, otherwise *always* "native-like" appropriate complete sentences.

C. *Amount of Communication*
   1. *Virtually no* relevant information was conveyed by the student.
   2. *Very little* relevant information was conveyed by the student.
   3. *Some* relevant information was conveyed by the student.
   4. *A fair amount* of relevant information was conveyed by the student.
   5. *Most* relevant information was conveyed by the student.
   6. *All* relevant information was conveyed by the student.

D. *Effort to Communicate*
   1. Student withdraws into long periods of silence, without any apparent effort to complete the task.
   2. Student makes *little* effort to communicate, what he does do is "half-hearted," without any enthusiasm.
   3. Student makes *some* effort to communicate, but still shows a rather "disinterested" attitude.

4. Student makes an effort to communicate but does not use any non-verbal resources, such as gestures.
5. Student makes a real effort to communicate and uses some non-verbal resources such as gestures.
6. Student makes a special (unusually high) effort to communicate and uses all possible resources, verbal and non-verbal, to express himself or herself.

APPENDIX E

The Daiei's English Proficiency Level Breakdown
(slightly revised)

(A) Level – International Executive Level
Fluent both in daily conversation and in specific situations, he can be precise in many fields. His high-level command of the language enables him to participate effectively, intelligently and logically in international conferences, to manage the overseas branch of a company, can conduct reliable negotiations, with perceptive understanding of Western thinking. Can fully comprehend Western speech and can convey the finer nuances as well as colloquial speech. May be competent enough to occasionally act as interpreter in social or formal situation. Can cope with study abroad, alone, e.g. post-graduate work, etc.

(B) Level – International Business Level
Competent and fluent enough to participate actively in conversations with Westerners, can discuss business without too many difficulties. May make some grammatical errors, minor or otherwise, but no serious problems in vocabulary, sentence structures, patterns, pronunciaion, etc., in communicating. Can express himself relatively well and can be considered for job abroad. Though some lack of self-confidence may be shown, he should be able to handle most unexpected problems social or job-wise, without due stress. His fluency may not be on the par of the Westerners', but his comprehension and perception is almost up-to-par. Can be considered and is willing for long-term study or job abroad.

(C) Level – Basic Business Level
Can handle uncomplicated business related conversations, capable of most daily conversations on a social level, but fluency is not up-to-par with comprehension. Can express himself, though with some difficulty. Responses may be a little slow due to his concern regarding his English ability, and may be slightly ill-at-ease with Western speech and/or Westerners. Can fulfill overseas job, but may not handle it alone, easily, on a long-term basis. May be able to handle himself linguistically

working as a trainee at an overseas branch. May be capable of auditing undergraduate work in an overseas university.

(D) Level – Basic Fluency Level
As he has already acquired basic vocabulary, sentence structures and patterns, he can generally understand daily conversation. Understands and is able to communicate only in very specific situations: meeting or assisting customers or guests, check-in and out of hotels, able to deal with people on a basic level, but not ready for more complicated discussions. Cannot convey complicated opinions and/or information. In a business conversation, can handle only on limited basis. May be able to fulfill his overseas duties if in a group on a short-term basis. Whether he is ready for working overseas may be questionable, and up to the discretion of his employer.

(E) Level – Basic Social Level
Can conduct simple social, basic conversation, greetings, etc. Can comprehend partially what the speaker says, but has difficulty in responding and expressing himself. His practical knowledge of advanced structures or vocabulary is lacking. If a lack of self-confidence is evident, in both himself, his job, etc., and his English, one can anticipate many problems, particularly if sent abroad alone only to be aggravated further by his too limited English.

(F) Level – Elementary Level
If the subjects are limited to simple, daily matters, he can converse, somewhat, though in a highly unskilled way and with frequent death pauses, and sometimes or often responds incorrectly. There are still deficiencies in fundamental sentence structures, vocabulary, etc. Can understand to some extent if spoken to slowly, repetitively, and using simple words.

(G) Level – Introductory
Very basic level in English conversation. Can understand simple phrases and sentences only, and if spoken to slowly and clearly. Though restricted to the above level, he can respond, though his response may be long overdue and in a one or two word answer. Exchange of opinions and/or ideas almost impossible. Acquisition of further, intensive study of basic English is a necessity. Real conversation ability at this level is almost nil.