Annual Review of Applied Linguistics (1998) **18**, 192–207. Printed in the USA.

ASSESSING SPEAKING

Jean Turner

AT AND BEYOND ORAL PROFICIENCY TESTING

The publication of the ACTFL (American Council on the Teaching of Foreign Languages) Guidelines (1986) and the creation and popularization of the ACTFL Oral Proficiency Interview (ACTFL-OPI) have had a profound effect on foreign and second language instruction and assessment, drawing attention to language students' abilities to use language in performing particular functions and tasks rather than to what they have learned about language. The growing interest in communicative language teaching, with its emphasis on meaningful interaction in the language as opposed to knowledge of linguistic rules, has complemented interest in the ACTFL Guidelines' descriptions of functional language ability and the interview-format oral proficiency interview. The movement toward communicative language teaching, research inspired by the ACTFL Guidelines, and the appeal of a face-to-face oral proficiency measure, initiated a decade of investigation of various features of oral proficiency interviews, including their reliability and construct validity (Dandonoli and Henning 1990, Henning 1992, Shohamy 1983), their comparability to other test formats, and their concurrent validity (Clark and Hooshmand 1992, Reed 1992, Shohamy 1994, Stansfield and Kenyon 1992). At the same time, researchers and educators—attempting to devise assessments that were congruent with the communicative approach that they adopted in their classrooms—explored various techniques for testing oral skills in specific educational settings and for specific programs of instruction (Day and Shapson 1987, Lindblad 1992, Manley 1995, St. John 1992).

This chapter first outlines the nature of the early research in interview-format oral proficiency testing, then reports on new directions in the investigation of the construct validity of interview-format tests and other types of oral skills tests through analyses of examinee, interviewer, and rater performance. Research in the development of empirically-derived rating scales is reviewed as well. The final

section discusses the findings of several studies that report on the development of oral skills tests for specific environments and the authors' efforts toward the integration of teaching and assessment. An on-going concern throughout the chapter is the validity of oral proficiency and oral skills tests, especially with regard to score use and generalizability.

EARLY RESEARCH IN INTERVIEW-FORMAT ORAL PROFICIENCY TESTING

In 1988, a special issue of *Studies in Second Language Acquisition*, addressing the assessment of foreign language proficiency, encapsulated the main issues and agendas at the time. Clark and Clifford (1988) described Interagency Language Roundtable (ILR)/ACTFL OPI procedures and scales and delineated a research agenda for investigating the reliability and validity of these interview-format oral proficiency tests. Bachman (1988) summarized his earlier criticisms of the ACTFL descriptors and scales as well as the ACTFL Oral Proficiency Interview (ACTFL-OPI) itself, noting that such standardization of oral interviews reduced random error and consequently was beneficial to reliability, but it introduced systematic error as a result of the test method itself, with a negative impact on the validity of oral proficiency interviews. Bachman reiterated the importance of investigations into the construct validity of the ACTFL-OPI (highlighting the notion of accountability), and the validity of the ACTFL-OPI for the various purposes for which it is used. He suggested a rigorous research agenda aimed at resolving the conceptual, reliability, and validity problems he claimed the ACTFL-OPI presented. Lantolf and Frawley (1988), in the same volume, had a much less optimistic view of interview-format tests, concluding "that the one place we cannot properly study what people do with or through language is the oral interview" (1988:191) because an interview, by its very nature, is unlike many non-test interactions, being directed by the interviewer rather than allowing shared responsibility for direction.

A special issue of *System* (1992), on the topic of oral proficiency testing, included research that continued the investigation of the construct validity of interview-format tests (Henning 1992, Reed 1992) and the concurrent validity of tape-mediated and videoconferencing-based interview-format tests (Clark and Hooshmand 1992, Stansfield and Kenyon 1992); however, this research did not directly address questions raised regarding the generalizability of performance on these tests or the nature of interview-format interactions. Other articles, by Lazaraton (1992) and Douglas and Selinker (1992), did explore these latter questions and took the investigation of the validity of interview-format tests in new directions through analyses of examinee, interviewer, and rater performance.

EXAMINEE, INTERVIEWER, AND RATER PERFORMANCE IN INTERVIEW-FORMAT TESTS

Despite the volume of research on the reliability and construct or concurrent validity of interview-format tests that followed the publication of the ACTFL Proficiency Guidelines and the ACTFL-OPI, remarkably, examinee and interviewer performance was not investigated. In 1989, van Lier asked two important questions that gave new direction and impetus to research on interview-format tests: First, does examinee performance on interview-format proficiency tests resemble non-interview discourse and second, should it?

Although it was never claimed by the producers of interview-format proficiency tests that they measure conversational skill, both the language teaching community and test users seem to assume that, because interview-format tests are direct (i.e., they measure oral skills by having the examinees actually speak), they must measure conversational ability as well as the ability to perform the tasks and functions that the procedures are designed to elicit. The generalizability of interview-format performance remains largely unexplored although a group of valuable and interesting studies on interview-format tests are beginning to offer some insights by describing examinee, interviewer, and rater discourse and behavior.

Lazaraton (1992) reported on a conversational analysis of twenty interview-format test performances. She described the phases that occur in a typical purposive interview and concluded that while examinee performance appears to follow the general structure of conversation, turn taking was different because of the interview format. She summarized her findings as follows: "The most that can be said about the question 'interview or conversation?' is that the encounters share features with conversations, but they are still characteristically instances of interviews" (1992:383).

Lazaraton (1996) used a similar conversational analysis approach to describe the types of interlocutor support that occur in the interview-format component of CASE, the Cambridge Assessment of Spoken English. Like Ross and Berwick (1992) and Ross (1992), she found that interviewers used many accommodations, or supportive practices, that occur in non-test, non-interview interactions between native speakers and non-native speakers, such as priming topics and slowed speech. However, Ross and Berwick, and Ross suggested that degree of accommodation could be an indicator of a rater's sense of an examinee's proficiency and might be investigated "to qualify the assessment of the interviewee's performance" (Ross 1992:183). The findings of Ross and Berwick, Ross, and Lazaraton, that some of the features of non-interview oral interactions are present in interview performance, are positive ones given that many test score users seem to believe that interview-format tests measure conversational skill. However, Lazaraton (1996) also found that the appearance of the facilitative

features she investigated was not consistent. As she noted, "this raises questions about its impact on candidate language use, and on the rating of that language" (1996:166). Her caution seems sensible as more information is assembled regarding examinee performance. Johnson (1997) and Johnson and Tyler (forthcoming) analyzed representative interview-format test performances in terms of the features of turn distribution and negotiation of topic, concluding that the interactions they examined did not resemble natural conversation with respect to these discourse features.

Young and Milanovic (1992), in their description of how task and other variables impact the nature of interview-format discourse, concluded that an asymmetrical contingency model holds for interview-format discourse, as was suggested by van Lier (1989), but that the measures of dominance that the authors investigated "do not capture the underlying control over the right to speak because...the candidate has the right to speak more than the examiner" (Young and Milanovic 1992:417). They reported that the oral proficiency interviews they examined did not resemble "the collaborating management of talk by both parties that we believe to be the structure of non-testing conversations" (1992:421). Young (1995) investigated the discourse features of intermediate versus advanced language learners and found that, although intermediate and advanced examinees differ on some tasks in terms of duration, pace, volume, and persistence of topic, some of the supposed characteristics of intermediate versus advanced learners represented in the rating scales were not substantiated in the actual performance of intermediate and advanced learners. He also reported that there were no significant differences in the features of interviewer discourse. He noted that this appears to contradict the conclusions of Ross (1992) and Ross and Berwick (1992), that accommodation by interviewers seems to reflect the interviewers' perceptions of the examinees' ability and that degree of accommodation might consequently be useful for qualifying examinee performance. These conflicting findings may simply reflect the complexity of performance on interview-format tests and support these authors in their call for further investigation into examinee and interviewer performance so that an empirical foundation can be generated for the interpretations that so many users make of interview-format test performance.

The first of van Lier's (1989) critical questions, whether examinee performance on interview-format tests resembles conversation, has now undergone some investigation. Although a clear answer has not been revealed, the studies discussed above have defined useful methodologies and questions for further research in this important area. The second question posed by van Lier, whether interview-format performance should resemble conversation has not been investigated, and in fact, it might be the more important of the two questions since it underlies the generalizability and validity of many test score users' interpretations of interview-format test scores.

FACTORS THAT AFFECT EXAMINEE AND RATER PERFORMANCE

Although test developers and test score users hope that the primary factor affecting test performance is examinee ability, it is recognized that many other variables, including features of the test, the environment, and the participants, also have an impact on an individual's test performance. Bachman's (1990) discussion of test method facets summarized and categorized some of these variables. He also called for a research program that would explore the impact of these variables on test scores and performance.

Douglas and Selinker (1992) investigated the impact of a test's content-specificity on scores and performance through analysis of examinee performance on two tests; a tape-mediated test of general speaking ability (the SPEAK, developed by Educational Testing Service) and a test developed at Iowa State University that parallels the SPEAK in structure and format but is discipline specific, with chemistry as its content. Douglas and Selinker used a third test as a point of reference for their comparison of examinees' performance on the general test and the discipline-specific test, a test developed by Iowa State to test international teaching assistants' communicative ability. They found that the discipline-specific test was more difficult than the general one for the 31 examinees whose performance they analyzed, even for the 22 whose major was chemistry. They also discovered that for the majority of the examinees, the discipline-specific test produced more complex discourse than the general one. The discipline-specific test was slightly more difficult for the raters to score consistently; however, performance on the discipline-specific test correlated more consistently with recommendations for type of teaching assignment than did performance on the general speaking test.

Douglas (1994), recognizing that raters' interpretations of scales cannot be completely controlled, expanded the investigation of factors that impact test performance by looking at rater behavior. In this study, the test performances of six examinees on a tape-mediated test designed to measure the speaking proficiency of students in an agriculture program were analyzed quantitatively and qualitatively. The quantitative analysis of the six allowed them to be paired on the basis of their grammar and comprehensibility, vocabulary, fluency, and organization scores. The qualitative analysis supported the author's hypothesis that even similar quantitative scores "represent qualitatively different speaker performances" (1994:133), with differences identified in degree of accuracy, length and complexity of responses, and precision of vocabulary. Although these findings could be due to generally poor rating or individual rater's bias to certain features of examinee performance, Douglas noted that the raters were influenced by features of test performance that are not included in the scoring rubric. He also noted that, "if we wish to use raters' judgments about learner performance as evidence of underlying language ability, we need to understand more thoroughly the bases upon which the raters make those judgments" (1994:135). An empirically-based

approach to the development of ratings scales, as is discussed in the following section, may prove useful in these investigations.

Wigglesworth (1997) investigated the impact of planning time on test performance using a tape-mediated test called *ACCESS:* (Australian assessment of communicative English skills). Planning time was manipulated as a counter-balanced variable in four sections of the test, with planning time provided in two of the sections and not provided in the other two. Although there were no significant differences in the scores of the performances with planning time and without planning time, Wigglesworth reported that planning time seemed to interact with examinee proficiency level and task difficulty. Higher ability examinees seemed to benefit from planning time on some tasks, especially more difficult ones, while planning time did not seem to benefit examinees at lower levels of ability except on the general discussion task. However, as Douglas (1994) reported, individuals with similar scores showed discourse-level differences in performance.

Fulcher (1996b) reported on an investigation of the number of interlocutors as a factor impacting oral skills test scores and performance. Forty-seven examinees completed three tasks. Two were one-to-one interactions between the examinee and the examiner, one based on a picture prompt and the other based on a reading prepared just before the interaction. The third task involved a prepared group discussion on a particular topic. Fulcher found that, in general, the examinees preferred the group task. Most reported that it seemed like a more natural interaction to them than the one-to-one tasks, that it provoked less anxiety, that it and the text-based one-to-one interaction were better measures of their ability than the picture-based task, and that the level of difficulty of the task seemed reasonable. Fulcher conducted G-study and Rasch analyses of the scores to address the generalizability of the tasks and reported that generalization seems possible when the scale used is not based on test method:

> The data and argument presented here suggest that, although there is a task effect in oral testing which has frequently been commented on in the literature, this may not be as large as has often been assumed when rating scales which do not contain descriptors which refer to test method facets are used. That is, large task effects may be an artifact of the rating scale used (1996b:37).

Perhaps scales which are referenced to type of task without an *a priori* empirical basis for doing so obscure critical differences in performance. This conclusion might explain Douglas's (1994) and Wiggleworth's (1997) findings of no significant score differences despite discourse level differences among examinees who completed tests varying in content-specificity or availability of planning time.

The impact of rater background variables has been investigated by Brown (1995), Lumley and McNamara (1995), and Chalhoub-Deville (1995a; 1995b).

Brown examined rater performance on an interview-format test of Japanese for tour guide trainees. Of the 33 raters, some had experience as tour guides, some as language teachers, and some in both professions. Using a multifaceted Rasch analysis approach, Brown discovered that, although there were no significant differences in the different types of raters' rankings, the raters did seem to differ in terms of their perceptions of the operation of the scale and in the harshness of their perceptions of the criteria. The scale that the raters used was developed through collaboration among the different types of raters. On the basis of her findings, Brown suggested that "if each group were to develop its own assessment framework..., they may, in fact, through the inclusion or weighting of specific criteria, produce schemes which lead to quite different evaluations of candidates' ability" (1995:13). She posed a thought-provoking question: Given that language experts, professional experts, and people with whom the examinees will interact when performing their job differ in their perceptions of examinee performance and the criteria they employ when assessing performance, who should participate in devising rating scales for career-related performance tests?

Lumley and McNamara's (1995) study contributed to the issue of rater performance in two ways, first as an investigation of the usefulness of multifaceted Rasch analysis (FACETS) for scoring, and second as an examination of the stability of rater characteristics over time, an important issue for training as well as score reporting. Lumley and McNamara's premise was that, if patterns in rater behavior can be identified, then error from these sources can be compensated for, perhaps through systematic adjustment of raw scores. Elimination of differences among raters might not always be the goal (cf. Moss 1994); however, as the authors noted, "differences and similarities [among raters] need to be taken into account in determining the best estimate of candidates' abilities at the time of the analysis of data from actual test administrations" (Lumley and McNamara 1995:59). Using ratings of the Occupation English Test conducted across a duration of 20 months, the authors discovered that there was variation in degree of harshness across time and raters. However, they also discovered that with appropriate anchoring, it was possible to identify raters who showed significant variation despite training. In their words, "this appears clear enough justification for using FACETS analysis of performance test data where no more than two raters are involved in assessing each candidate, since it is able to take relative severity of judges into account and make adjustments to estimates of candidate ability" (1995:69).

Chalhoub-Deville (1995b) investigated rater performance in the context of rater behavior's impact on construct validity. She noted that the "fundamental issue in construct validation is to uncover the attributes of the construct underlying test scores" and that "performance-based L2 oral test scores summarize, in addition to learners' ability, construct irrelevant variance from sources such as the task and the rater" (1995b:17). Shohamy, Gordon, and Kraemer (1992) and Hadden (1991) among others have reported that both task and rater have an impact

on ratings; Chalhoub-Deville added to this knowledge by attempting to identify the dimensions of task and rater which might be related to variation in examinee and rater performance. Using language samples that were generated by six learners of Arabic, two-minute segments from each of three tasks (interview, narration, and read-aloud) were compiled and the 18 speech samples were rated by 82 raters. The raters represented three different profiles: teachers of Arabic who were native speakers of Arabic and living in the U.S., non-teachers (university students) who were native-speakers of Arabic and living in the U.S., and non-teachers (university students) who were native-speakers of Arabic and living in Lebanon. Using a multidimensional scaling procedure, Chalhoub-Deville identified three dimensions, *grammar-pronunciation*, *creativity in presenting information*, and *amount of detail,* all of which varied in interpretation and weight across the tasks and raters. The author echoed Brown's (1995) question; given the differences among raters with different profiles, who should devise rating scales and who should rate? The difference among tasks in her work, which parallels findings by others, supports the need for different types of tasks in oral skills testing.

Although she did not answer the question of who should devise rating scales and who should rate, Chalhoub-Deville (1995a) recommended developing scales based on empirical evidence of patterns in examinee and rater behavior. Because task and background variables seemed to have an impact on the schemes that raters employ, she recommended that researchers who are investigating L2 oral proficiency use "a variety of tasks, which would provide adequate portrayal of the construct [L2 proficiency] and...employ diverse audiences—that is, criterion judges" so that "any systematic relationship between learners' L2 proficiency, the task, and the rater can be detected and described, leading to better understanding of the proficiency construct" (1995a:275).

The investigations to date into the nature of examinee, interviewer, and rater performance and variables that affect their performance are revealing but inconclusive. They offer new perspectives on test validity, further evidence of the variability of test performance, and insights into the impact of test method and rater variables on test performance and test scores. Important questions regarding who should be involved in the development of rating scales have been posed, and general suggestions for more effective approaches to scale development have been made.

DEVELOPING RATING CRITERIA FROM EMPIRICAL EVIDENCE

The notion of using rating scales to help train raters and to establish reliability and validity are generally accepted; however, questions have arisen regarding how these scales should be created. They are now usually created by language or testing experts with 'validation' taking place after the fact (if at all). One of the main criticisms of the ACTFL Guidelines and the ACTFL-OPI is the lack of an empirical basis for the descriptors that form the foundation, although

some studies have attempted to validate the descriptors after the fact (Stansfield and Kenyon 1996). Fulcher (1996a) proposed an approach to scale development which would be *a priori* and empirically-based. Focusing on the notion of fluency, he used Grounded Theory to derive eight features of fluency from his data (the interview-format test performances of 21 Greek-speaking learners of English); he then used discriminate analysis to explore the coincidence of frequency counts of these features with placement on the English Language Testing System (ELTS) scale. He found that, with one exception, the grouping by fluency features corresponded with ELTS placement. From these results, he concluded that such empirically-based scales could serve as the basis for more specific, more informative scales that could be related to actual test performance and consequently allow more conclusive validation than scales based on expert opinion or theory. This approach might actually capture the elements of language performance that are apparent in analyses of examinees' performance but that do not affect scores, as were reported by Douglas (1994) and Wigglesworth (1997). Fulcher noted, "Until test researchers and developers take seriously the validity of tests at the development phase rather than as a *post hoc* notion, the problem of the indeterminacy of validation studies and the uninterpretability of test scores will remain serious" (1996a:228). He recommended that research into the empirical derivation of rating scale criteria "should be carried out into the description and operationalization of constructs for language testing, reinforcing the necessary link between applied linguistics, second language acquisition research and language testing theory and practice" (1996a:228). His recommendation offers a useful direction for research in rating scale development and the validity of oral skills tests.

Upshur and Turner (1995) proposed another empirical approach to building rating scales, one that is suitable for classroom-based testing. The authors claimed that this approach could address the limitations of existing proficiency scales in settings where teachers assess the progress of their students in communicatively-oriented language classes. These limitations include the fact that existing proficiency scales tend to have rather broad, imprecise descriptors; they consequently do not reflect teachers' more narrowly defined objectives closely enough to provide meaningful measurement of student learning. Another limitation of the existing scales for classroom testing is that they are based on descriptions of features that may not co-occur in actual student performance. As with all scales, the usefulness of the existing proficiency scales is limited by raters' individual interpretations of the descriptors; however, an additional, related issue is that teachers' standards may shift from the beginning to the end of a course or differ from one course to another.

Through the approach proposed by Upshur and Turner, instead of attempting to match test performance to verbal descriptors from scales, the scale is derived from a hierarchical set of binary questions formed by teachers using a subset of examinee performances. These scales are developed for particular tasks

and are intended to show how well examinees perform only on these types of tasks; score interpretations are not intended to be generalizable but when carefully developed, the scales have the potential to assess students' learning more accurately than the less focused proficiency scales. Upshur and Turner called scales derived through this approach EBB scales—empirically-derived, binary-choice, boundary-definition scales. To develop the EBB that illustrated their discussion, a teaching team worked with actual student performance samples, sorting them impressionistically into a high group and a low group. The team then refined the sort and formed a question that captured the most salient, critical characteristic that distinguished the two groups. Then, within the high and low groups, the papers were sorted, and critical questions were formed that distinguished among levels within the high and low groups. The scale was created through consensus among the teachers on the nature of the defining characteristics. The authors claimed that an EBB-scale approach addresses the feasibility, reliability, and validity limitations of existing proficiency scales for use in specific educational contexts by describing boundaries that are concrete, simple, and precise. Another useful feature is that, because the scales are empirically derived through consensus among teachers, there may be a beneficial impact on instruction as the teachers come to agreement in defining critical characteristics. The approach merits further investigation because of its feasibility, its greater appropriateness for measuring student progress, and the benefits for teachers and language programs.

TEST DEVELOPMENT PRINCIPLES AND THE DEVELOPMENT OF PROGRAM-SPECIFIC TESTS

As Upshur and Turner (1995) noted, when teachers invest in a communicative approach to language teaching, a need grows for measures of oral skills that can be used by teachers in classrooms. This may be especially true when working with children for whom existing oral skills tests are frequently inappropriate. Butler and Stevens (in press) set forth principles for assessing the oral skills of K-6 students, stressing the relationships among teaching, learning, and assessment. While focusing on oral skills assessment, the discussion is intended as a model for assessment of communicative skills in general. The authors encourage frequent, on-going assessment and the development of student profiles. O'Malley and Pierce's (1996) description of oral portfolios is offered as an example of how to assemble the information from on-going assessment with both individual and group tasks included. Evaluation criteria are discussed in the context of examples from several educational settings, including book talks and a group discussion task. Perhaps most importantly, clear principles for developing assessments for classroom use are defined, including the use of multiple measures and the importance of establishing developmentally-appropriate, performance-based criteria that examinees are aware of and understand.

Carpenter, Fujii, and Kataoka (1995) described the development of an oral interview for children between the ages of 5 and 10 who participated in an immersion program in Japanese, noting that immersion education has moved beyond a point where the primary concern of educators and parents is whether content learning is acceptable to a point where there is interest in language learning. This latter goal requires tests that indicate levels of participants' language learning. The authors also noted that changes have been made in immersion programs to address issues such as the development of classroom dialects. They maintained that these changes should be made on the basis of a good understanding of language learning in immersion setting, therefore necessitating the "careful analysis of the acquisition process as it occurs daily in the classroom, as well as children's progress over shorter timespans" (1995:159). They claimed that existing language tests for children (none of which are available in Japanese) tend to reflect the original content learning concerns of educators and parents, or measure grammar and reading comprehension; consequently, they cannot serve the function of reporting on communicative language learning. Acccording to the authors, these tests may even penalize children for pragmatic sophistication because their formats tend to encourage short responses.

The authors' goal was to develop a test that would be fun for children, pragmatically appropriate, and capable of measuring a range of content and speech styles, while yielding assessments of ability that are comparable across children and programs. The test they created included six subtests, each intended to elicit a different kind of ability and with enough variety that different children's personalities and levels of ability might be accommodated. The test was piloted with 40 children and the performances of two children, a first grader and a fourth grader, were analyzed through comparison with their performance on a Japanese translation of the Spanish Oral Proficiency Test. The authors found that performance on their test resulted in longer responses with a higher proportion of complete sentences and more and greater variety of verb and noun use. There was also an indication that a built-in feature of the test, which allows for a variety of questions and materials, elicited a "representative sample of children's abilities, independently of their shyness or talkativeness" (1995:169). The test also prevented boredom and a possible impact of test familiarity in the event the test had to be taken by an individual on several occasions. They claimed that the assessment criteria used were independent of topic and vocabulary and were based on "activities that are widely accepted as good for language teaching as well as language testing" (1995:172), thus reflecting several of the principles that Butler and Stevens (in press) described for developing useful tests of children's oral skills. Unfortunately, Carpenter, *et al.* did not fully explain the rating criteria. However, the authors noted that designing a rating scale and piloting the test with children who are both native speakers and non-native speakers of Japanese is on their test development agenda. Despite the lack of detail in the description of the rating criteria and procedure, the project is valuable in the solutions it offers to problems that arise when designing tests to measure children's learning.

Haggstom (1994) described another solution to measuring students' learning. Her video-based testing procedure was based on the assumption that when teaching for communicative competence, tests should entail communicative tasks. She also noted that tests should allow measurement of students' progress and mastery of specific material as well as being accurately scoreable and feasible in terms of administration and grading time. The testing method she described employed video-taping of students' performances in communicative activities that were typical of her classroom activities. Because the testing was integrated with classroom activities, on-going frequent assessment was possible. In this 50-minute test procedure, the teacher moved around the room with a camcorder and videotaped each student on three occasions as he or she participated in a small group activity. Two example activities were described and criteria for selecting tasks and scoring responses were provided. Although this approach might not be feasible or appropriate in all classrooms, the description offers useful insights into integrating instruction and assessment and designing tests for specific learning environments.

Manley (1995) reported on one school district's commitment to familiarizing foreign language teachers with oral skills testing, and to developing and giving tape-mediated oral skills tests for French, German, Spanish, and Japanese. The 30-month project included as its first step an overview of oral proficiency testing and workshops in which teachers explored major issues in oral skills assessment. Teachers who volunteered and were chosen to be test writers then identified areas and topics to be tested and optimal formats for testing. They also developed an assessment grid to be used for scoring the tests, after which the tests were piloted and the insights gained through that piloting shared and incorporated into the plan for the following year's activities. The tests developed in the second year were also piloted and reviewed to inform the test writers and other teachers about the progress that had been made and what had been learned. During a summer workshop, the data were compiled and test revisions were made. The author reported that the benefits of the project included the teachers' gaining personal experience in the development of oral skills tests as well as currency in important developments in language education. The author also reported that the project had a positive effect on the morale of the teachers, bringing them together to work through complex problems and come to consensus. Manley's study represents some of the beneficial outcomes of the current trends in the development of oral assessments and scales—involvement of teachers, integration of learning and assessment, and the development of clear, understandable rating criteria.

Butler and Stevens (in press) remind us of important principles for creating an assessment program that integrates testing and teaching. The work by Carpenter *et. al.* (1995), Haggstrom (1994), and Manley (1995) demonstrates a range of solutions that practicing teachers and test developers have adopted as they create appropriate assessments for their students' progress. As is typical of tests developed for specific programs and purposes, the instruments may not be

appropriate in other educational settings, but the reports model not only how to approach integrating teaching and assessment but also the benefits of doing so.

One of the principles that Butler and Stevens proposed is that criteria should be performance-based and clear to examinees. The development of such criteria may be informed by Manley's (1995) description of the collaborative test development effort conducted in her school district and by continued investigation of empirically-based rating scales (Fulcher 1996a, Upshur and Turner 1995). Sounder scales will certainly result in more valid measurement. Investigations into the impact of features of the test, environment, and participants on test performance and scores provide additional information and new techniques for identifying the sources of error, and perhaps accounting for them in score reporting (Lumley and McNamara 1995).

The concerns regarding the nature of interview-format discourse expressed by Lantolf and Frawley (1988) and summarized by van Lier's (1989) questions—Do interview-format test interactions resemble conversation? Should they?—are addressed through on-going research on the features and discourse structure of interview-format test interactions (Johnson 1997, Johnson and Tyler forthcoming, Lazaraton 1992, Ross 1992, Ross and Berwick 1992, Young 1995, Young and Milanovic 1992). The findings are contradictory and inconclusive, but work in this area continues to inform the issues of the validity and generalizability of interview-format tests while the gains made through this decade of research allow us to move toward more meaningful measurement of oral skills.

UNANNOTATED BIBLIOGRAPHY

American Council on the Teaching of Foreign Languages. 1986. *ACTFL proficiency guidelines*. New York: American Council on the Teaching of Foreign Languages.

Bachman, L. 1988. Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition*. 10.149–164.

_____ 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Brown, A. 1995. The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*. 12.1–15.

Butler, F. A. and R. Stevens. In press. Oral language assessment in the classroom. In *New directions in student assessment*. 214–219. [Special issue of *Theory into Practice*. 36.4.]

Carpenter, K., N. Fujii, and H. Kataoka. 1995. An oral interview procedure for assessing second language abilities in children. *Language Testing*. 12.157–181.

Chalhoub-Deville, M. 1995a. A contextualized approach to describing oral language proficiency. *Language Learning*. 45.251–281.

_____ 1995b. Deriving oral assessment scales across different tests and rater groups. *Language Testing*. 12.16–33.

Clark, J. L. D. and R. T. Clifford. 1988. The FSI/ILR/ACTFL proficiency scales and testing techniques: Development, current status, and needed research. *Studies in Second Language Acquisition*. 10.121–147.

_____ and D. Hooshmand. 1992. "Screen-to-screen" testing: An exploratory study of oral proficiency interviewing using video teleconferencing. *System*. 20.293–304.

Coniam, D. 1995. Towards a common ability scale for Hong Kong English secondary-school forms. *Language Testing*. 12.182–193.

Courtney, M. 1996. Talking to learn: Selecting and using peer group oral tasks. *ELT Journal*. 50.318–326.

Dandonoli, P. and G. Henning. 1990. An investigation of the construct validity of the ACTFL oral proficiency guidelines and oral interview procedure. *Foreign Language Annals*. 23.11–22.

Day, E. M. and S. Shapson. 1987. Assessment of oral communicative skills in early French immersion programmes. *Journal of Multilingual and Multicultural Development*. 8.237–260.

Douglas, D. 1994. Quantity and quality in speaking test performance. *Language Testing*. 11.125–143.

_____ and L. Selinker. 1992. Analyzing oral proficiency test performance in general and specific purpose contexts. *System*. 20.317–328.

Edwards, A. L. 1996. Reading proficiency assessment and the ILR/ACTFL text typology: A reevaluation. *Modern Language Journal*. 80.350–361.

Fulcher, G. 1987. Tests of oral performance: The need for data-based criteria. *ELT Journal*. 41.287–291.

_____ 1995. Variable competence in second language acquisition: A problem for research methodology? *System*. 23.25–33.

_____ 1996a. Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*. 13.208–238.

_____ 1996b. Testing tasks: Issues in task design and the group oral. *Language Testing*. 13.23–52.

Hadden, B. 1991. Teacher and nonteacher perceptions of second-language communication. *Language Learning*. 41.1–24.

Haggstrom, M. 1994. Using a videocamera and task-based activities to make classroom oral testing a more realistic communicative experience. *Foreign Language Annals*. 27.161–175.

Halleck, G. B. 1992. The oral proficiency interview: Discrete point test or measure of communicative language ability? *Foreign Language Annals*. 25.227–231.

Harlow, L. L. and R. Caminero. 1990. Oral testing of beginning language students at large universities: Is it worth the trouble? *Foreign Language Annals.* 23.489–501.

Henning, G. 1992. The ACTFL Oral Proficiency Interview: Validity evidence. *System.* 20.365–372.

James, R. 1996. CALL and the speaking skill. *System.* 24.15–21.

Johnson, M. 1997. What kind of a speech event is the oral proficiency interview? Problems of construct validity. Washington, DC: Georgetown University. Unpublished manuscript.

_____ and A. Tyler. Forthcoming. Re-analyzing the OPI: How much does it look like natural conversation? In R. Young and W. Y. Hae (eds.) *Talking and testing: Discourse approaches to the assessment of oral proficiency.* Amsterdam: John Benjamin.

Lantolf, J. P. and W. Frawley. 1988. Proficiency: Understanding the construct. *Studies in Second Language Acquisition.* 10.181–195.

Lazaraton, A. 1992. The structural organization of a language interview: A conversation analytic approach. *System.* 20.373–386.

_____ 1996. Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing.* 13.151–172.

Lindblad, T. 1992. Oral tests in Swedish schools: A five-year experiment. *System.* 20.279–292.

Lumley, T. and T. F. McNamara. 1995. Rater characteristics and rater bias: Implications for training. *Language Testing.* 12.54–71.

Manley, J. H. 1995. Assessing oral language: One school district's response. *Foreign Language Annals.* 28.93–102.

Matthews, M. 1990. The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *ELT Journal.* 44.117–121.

Moss, P. A. 1994. Can there be validity without reliability? *Educational Researcher.* 23.4–12.

Nibungco, J. T. and M. D. Williams. 1996. Designing oral assessment for nontraditional ESL students in a community college. *College ESL.* 6.85–94.

O'Loughlin, K. 1995. Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing.* 12.217–237.

O'Malley, J. M. and L. V. Pierce. 1996. *Authentic assessment for English language learners: Practical approaches for teachers.* Reading, MA: Addison-Wesley.

Prodromou, L. 1995. The backwash effect: From testing to teaching. *ELT Journal.* 49.13–25.

Raffaldini, T. 1988. The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition.* 10.197–216.

Reed, D. J. 1992. The relationship between criterion-based levels of oral proficiency and norm-referenced scores of general proficiency in English as a second language. *System.* 20.329–345.

Ross, S. 1992. Accommodative questions in oral proficiency interviews. *Language Testing*. 9.173–186.

_____ and R. Berwick. 1992. The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquistion*. 14.159–176.

Shohamy, E. 1983. The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning*. 33.527–540.

_____ 1988. A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition*. 10.165–180.

_____ 1994. The validity of direct versus semi-direct oral tests. *Language Testing*. 11.99–123.

_____, C. M. Gordon and R. Kramer. 1992. The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*. 76.27–33.

Stansfield, C. W. and D. M. Kenyon. 1992. Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*. 20.347–366.

_____ 1996. Comparing the scaling of speaking tasks by language teachers and by the ACTFL Guidelines. In A. Cumming and R. Berwick (eds.) *Validation in language testing*. Clevedon, Avon: Multilingual Matters. 124–153.

St. John, J. 1992. The Ontario Test of ESL Oral Interaction. *System*. 20.305–316.

Taylor, R. E. 1995. Assessing oral communication skills—reflections of an oral examiner. *World Englishes*. 15.131–136.

Thompson, I. 1995. A study of interrater reliability of the ACTFL oral proficiency interview in five European languages: Data from ESL, French, German, Russian, and Spanish. *Foreign Language Annals*. 28.407–422.

Upshur, J. A. and C. E. Turner. 1995. Constructing rating scales for second language tests. *ELT Journal*. 49.3–12.

Van Lier, L. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*. 23.469–508.

Wigglesworth, G. 1997. An investigation of planning time and proficiency level on oral test discourse. *Language Testing*. 14.85–106.

Young, R. 1995. Conversational styles in language proficiency interviews. *Language Learning*. 45.3–42.

_____ and M. Milanovic. 1992. Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*. 14.403–424.