A REVIEW AND CRITIQUE OF PROCEDURES FOR ASSESSING SPEAKING AND LISTENING SKILLS AMONG PRESCHOOL THROUGH GRADE TWELVE STUDENTS.

Donald L. Rubin
John Daly
James C. McCroskey
Nancy A. Mead

Language arts educators frequently regard large-scale assessment programs with suspicion. This antipathy is justified when testing does not support instruction, but rather undermines it. Teachers and students may be exploited, for example, by programs which feed highly aggregated data to administrators, but withold pedagogically useful feedback from instructors. Or instruments may tap readily measurable skills and knowledge, to the neglect of more appropriate learning outcomes which may be more resistant to well established measurement techniques.

Because large-scale assessment programs can threaten educational progress in these ways, the measurement enterprise demands the informed concern of educators. On the other hand, well motivated and well administered assessments may exert desirable effects on instruction. In the context of program evaluation, for example, tests of student achievement can serve a formative function, enabling teachers to "fine tune" their techniques. Examinations which are closely tied to educational practice can contribute to individualization by informing decisions which place students at appropriate points in instructional sequences. In addition, testing is a powerful "top-down" force for engineering curricular innovations. When those innovations are sound, assessment procedures can be said to exhibit pedagogical validity (Rubin, 1980).

Several factors render increased sensitivity to matters of educational measurement especially timely for educators in the field of speech communication. In conjunction with trends toward competency-based education, several state and local jurisdictions have already initiated large scale assessments of speaking and listening proficiency among public school students, and other jurisdictions are planning to do so (Backlund, 1981). Due, in part, to vigorous efforts by communication educators, Federal Basic Skills legislation now includes speaking and listening among its enumeration of targeted skills. While those projects directly enabled by the Basic Skills mandate comprise an immediate "market" for speaking and listening tests, the impact of this recognition of oral communication is likely to diffuse, and thus create additional demand for related assessment procedures.

Demand alone, of course, can not justify efforts at test development. Where large scale programs for assessing oral communication skills have been implemented, however, experience suggests that positive effects on instruction may accrue. The British Certificate of Secondary Education (CSE) examination of oral proficiency is

Donald L. Rubin is an Assistant Professor in the Departments of Speech Communication and Language Education at the University of Georgia; John Daly is Assistant Professor of Speech Communication, The University of Texas at Austin; James C. McCroskey is Chairman of the Speech Communication Department, West Virginia University; and Nancy A. Mead is associated with the Education Commission of the States.

COMMUNICATION EDUCATION, Volume 31, October 1982

a case in point. Established since the mid-Sixties, the CSE examination is available to students in a variety of forms (Hitchman, 1966, 1968; School's Council, 1966). British educators were aware that such tests could have potent "washback" effects on instruction, particularly in the absence of well accepted speech curricula (Wilkinson, 1968). Retrospective analyses of the CSE examination suggest that this assessment program has legitimized and guided classroom instruction in speaking and listening (Barnes, 1980; Wade, 1978).

Based on the premise, then, that large scale assessment can be a constructive force in oral communication instruction, the goals of this report are to (1) provide informational support for test utilization decisions and (2) suggest appropriate directions for further research and development in assessment instrument construction.[1] Several related documents preceded this investigation. Larson and his colleagues (1978) expounded a notion of functional communication competence and reviewed a number of measurement instruments. They did not, however, distinguish descriptive measures suitable for research from evaluative instruments designed for large scale testing of public school-aged populations. A project sponsored by the Massachusetts Department of Education (Brown, Backlund, Gurry & Jandt, 1979) compiled information about speaking and listening assessment instruments and found none which served its needs. Similarly, a team under the auspices of the Alberta Minister's Advisory Committee on Student Achievement (Plattor, Unruh, Muir & Loose, 1978) was unable to locate an appropriate instrument. The Speech Communication Association has endorsed a set of guidelines (originally developed by the Massachusetts project) for evaluating the adequacy of speaking and listening assessments (Criteria for Evaluating Instruments and Procedures for Assessing Speaking and Listening," 1979).

Survey of Available Instruments

In order to establish an information base of currently available measures of speaking and listening proficiency, the authors searched a variety of sources. These included standard reference guides to tests and assessment procedures (Buros, 1978; Johnson, 1976) as well as more specialized guides pertaining to language arts (Fagan, Cooper & Jensen, 1975; Grommon, 1976) and previous reviews of oral communication instruments (Brown, *et al.*, 1979, Larson, *et al.*, 1978; Plattor, *et al.*, 1978). Letters of inquiry were sent to selected commercial test publishers and to state and local educational agencies. The literature on foreign language testing was also consulted (e.g., Lange & Clifford, 1980; Richard, 1981). A search of ERIC system documents was conducted. Requests for information appeared in several professional newsletters.

Instruments which were ultimately selected for examination and review were those which at least nominally tested elements of speaking and listening skill, which were designed to yield evaluative judgments, and which were appropriate for preschool through grade 12 populations. Excluded, for example, were patent tests of general verbal ability which used oral language only incidentally. Measures which assumed a descriptive, research orientation were also eliminated from consideration. The final sample of instruments and procedures is described in Table 1. (Subsequent references to instruments are by numbers appearing in Table 1) Specimen copies of examinations, technical manuals, or other related documents were examined for each test included in the compilation.[2]

TABLE 1

CATALOGUE OF INSTRUMENTS REVIEWED

Instrument Number	Title	Source	Skills Tested	Target Populations	Mode of Administration
1	Brown-Carlson Listening Test	Harcourt, Brace & World, New York 10017	listening	high school, adult	administered orally and completed on standardized forms
2	California Achievement Tests: Listening	CTB/McGraw-Hill, Del Monte Research Park, Monterey, CA 93940	listening	primary	group administered; multiple choice, paper and pencil format
3	Circus Listen to the Story (Versions B, C, D)	Addison-Wesley, Reading, MA 01867	listening	k-3	group administered with multiple choice responses
4	Circus Say and Tell	Addison-Wesley Reading, MA 01867	oral language	pre-k to 3	individual, oral; child responds to a variety of stimuli
5	Cloze Listening Test	John S. Bowdidge, 2017 S. Oak Grove Avenue, Springfield, MO 65804	listening	secondary level	group administered; fill in the blank, paper and pencil format; tape recorded stimulus
6	Comprehensive Tests of Basic Skills	CTB/McGraw Hill, Del Monte Research Park Monterey, CA 93940	listening; visual decoding; auditory discrimination	early elementary	group administered; paper and pencil, multiple choice format
7	Communicative Evaluation Chart from Infancy to Five Years	Educators Publishing Service, Cambridge, MA 02138	oral language; listening; social development; auditory perception	infancy-5 years	observer records presence or absence of skills on basis of extended observation
8	Durrell Listening-Reading Series	Harcourt Brace Jovanovich, Inc., New York, NY 10017	listening	grades 1–9	group administered; multiple choice, paper and pencil format
9	Dyadic Task-Oriented Communication	C. A. Findley, ERIC Document Reproduction Service No. 145 629	speaking	elementary level	administered to pairs of students, one presents task, other responds; responses tape recorded
10	DYCOMM	B.H. Byers, *DYCOMM*: Dyadic Communication. Honolulu: University of Hawaii, 1973	speaking; listening; interaction	adaptable k-12	groups of 10 or more students work in dyads rotating among partners and tasks
11	Fullerton Language Test for Adolescents	Consulting Psychologists Press, Palo Alto, CA 94306	listening; auditory processing	11–18 years; learning disabled and nondisabled	individual administration

TABLE 1 (cont.)

Instrument Number	Title	Source	Skills Tested	Target Populations	Mode of Administration
12	Fundamental Achievement Series: Verbal	The Psychological Corporation, 757 Third Avenue, New York, NY 10017	listening; receptive language	grades 6–12	group administered; multiple choice, paper and pencil format; taped instruction
13	Gary, Indiana Oral Proficiency Examination	Gary Community School Corporation; Gary, IN 46401	speaking	grade 10	individual speech performance addressed to examiner
14	Glynn County Speech Proficiency Examination	CBE Demonstration Project, Glynn County, Board of Education, Brunswick, GA 31521	speaking	secondary level	simulated public hearing, students presenting arguments one at a time; responses videotaped.
15	Language Assessment Scales	Linguametrics Group, P.O. Box 454, Corte Madera, CA 94925	speaking; listening	grades 1–5; Spanish or English	multiple choice responses to oral presentations; oral imitation of sounds and words
16	Language Dominance Survey	Multilingual Center, Berkeley, California	speaking; listening	k–12; Spanish, English	individual administration
17	Language Facility Test	The Allington Corporation, 801 N. Pitt St., Alexandria, VA 22314	speaking	ages 3–15 for normal populations	individually administered; free responses to picture stimuli
18	Language Skills Communication Task	M. C. Wang, S. Rose, & J. Maxwell, *The Development of the Language Skills Communication Test.* Pittsburgh: University of Pittsburgh Learning Research and Development Center, 1973	speaking; listening; interaction	k–2	students work in dyads; responses are recorded for subsequent scoring
19	Listening Comprehension Tests	A. Wilkinson, L. Stratta and P. Dudley, *Listening Comprehension Tests.* Macmillan Education Ltd., Houndsmills, Basingstoke Hampshire England RG21 2X5	listening	ages 10–11, 13–14, and 17–18	group administered; paper and pencil; multiple choice format

TABLE 1 (cont.)

Instrument Number	Title	Source	Skills Tested	Target Populations	Mode of Administration
20	MACOSA Listening and Speaking Tests	E. Plattor, W.R. Unruh, L. Muir & K.D. Loose *Test Development for Assessing Achievement in Listening and Speaking* The Minister's Advisory Committee on Student Achievement Planning and Research Alberta Education, 10105 109 Street, Edmonton, Alberta, Candaa T5J 2V2	speaking; listening	grades 3, 6, 9, and 12	oral speaking test administered to small groups, each student responding in turn; responses tape recorded. written speaking test and listening test group administered, paper and pencil, multiple choice format
21	Massachusetts Assessment of Basic Skills Listening Test	Massachusetts Department of Education Bureau of Research and Assessment, Boston, MA 02116	listening	grade 12	group administered with tape recorded instructions, listening passages, and multiple choice response
22	Massachusetts Assessment of Basic Skills Speaking Test	Massachusetts Department of Education Bureau of Research and Assessment, Boston, MA 02116	speaking	grade 12	two-tiered system with classroom teachers rating typical speaking abilities, and individual interviews for students who fail to pass the initial screening
23	Measure of Communication Competence	S. C. Riccillo, *Children's Speeach and Communicative Competence,* Unpublished doctoral dissertation, University of Denver, 1974 University Microfilms No. 75-2210	speaking	ages 2½ to 4 years	individually administered, responses tape recorded
24	Metropolitan Achievement Tests: Listening Comprehension	The Psychological Corporation, 757 Third Avenue, New York, NY 10017	listening	k–9	group administered, multiple choice, paper and pencil format

TABLE 1 (cont.)

Instrument Number	Title	Source	Skills Tested	Target Populations	Mode of Administration
25	Michigan Educational Assessment Program Listening Test	Michigan Educational Assessment Program, Michigan Department of Education, P.O. Box 30008, Lansing, MI 48909	listening	grades 4, 7 and 10	group administered; paper and pencil, multiple choice format.
26	National Assessment of Educational Progress Pilot Test of Speaking and Listening	See, N.A. Mead, *The Development of an Instrument for Assessing Functional Communication Competence of Seventeen-Year-Olds.* Unpublished dissertation, University of Denver, 1977	speaking; listening; attitudes	age 17	group administered, multiple choice, paper and pencil format, tape recorded instructions
27	New York State Regents Comprehensive Examination in English, Listening Section	Division of Educational Testing, New York State Education Department Albany, NY 12234	listening	grade 12	group administered; examiner reads passages aloud; multiple choice format
28	New York Statewide Achievement Examination in English	Division of Educational Testing, New York State Education Department Albany, NY 12234	speaking; listening	grade 12	for speaking section, students present brief monologues on supplied topics in class; listening section is group administered; passages are read aloud; multiple choice format
29	Oliphant Tests: Auditory Synthesizing Test and Auditory Discrimination Memory	Educators Publishing Service; Cambridge, MA 02138	auditory memory	age 7–14	sounds are presented that examinee must hold in memory or discriminate
30	Oral Language Evaluation	EMC Corporation St. Paul, MN	speaking	elementary Spanish, English	individually adminstered, student's discussion of supplied stimuli is tape recorded and transcribed

TABLE 1 (cont.)

Instrument Number	Title	Source	Skills Tested	Target Populations	Mode of Administration
31	Profile of Nonverbal Sensitivity	R. Rosenthal, J.A. Hall, M.R. DiMatteo, P.L. Rogers, and D. Archer, *Sensitivity to Nonverbal Communication* Baltimore: John Hopkins University Press, 1979	nonverbal decoding	grades 3–6; high school	group administered; students view videotape or film; multiple choice response format
32	PRI Reading Systems (Oral Language Cluster)	McGraw-Hill, New York, NY 10036	listening	grades k–4	group administered with multiple choice answers
33	SRA Achievement Series, Levels A, B and C	Science Research Associates, Inc., 155 North Wacker Dr., Chicago, IL 60606	listening; auditory discrimination	grades k–3	group administered; paper and pencil, multiple choice format
34	Sequential Tests of Educational Progress (Listening)	Addison-Wesley, Reading, MA 01867	listening	grades 3–12	group administered using multiple choice responses
35	Stanford Achievement Test: Listening	Harcourt Brace Jovanovich, New York, NY 10017	listening	primary	group administered; multiple choice, paper and pencil format
36	Stanford Early School Achievement Test (Aural Comprehension)	Harcourt Brace Jovanovich, New York, NY 10017	listening	grades k–1	group administered; multiple choice format, paper and pencil
37	Situational Language Tasks	E. E. Conrad, R. K. Rentfrow, K. Meredith, and J. M. Fillerup, *Use of Situational Language Tasks in an Intra-TEEM and TEEM versus Comparison Evaluation.* Tucson, AZ: University of Arizona College of Education, 1976	speaking; listening; interaction	grades 1–3	includes whole-class discussion, and structured and unstructured small group discussion; talk is recorded and transcribed

TABLE 1 (cont.)

Instrument Number	Title	Source	Skills Tested	Target Populations	Mode of Administration
38	Speech in the Clasroom: Assessment Instruments	Bureau of Curriculum Services, Pennsylvania Department of Education, 333 Market Street, Harrisburg, PA 17126	speaking; speaking experience; attitudes	grades 1–6	assessment of speaking skills individually administered, others group administered; paper and pencil, multiple choice format
39	Test of Adolescent Language	PRO-ED, 333 Perry Brooks Building, Austin, TX 78701	speaking; listening	ages 11–18	speaking tests individually administered; listening tests group administered; paper and pencil, multiple choice format
40	Test of Listening Accuracy in Children	Communication Research Association, P.O. Box 11012, Salt Lake City, UT 84111	listening	grades k–6	group administered with examinee completing multiple choice questions
41	Torrance Tests of Creative Thinking (Oral Administration)	Scholastic Testing Service, Inc., 480 Meyer Rd. Bensenville, IL, 60106	creative thinking	grades k–3	individually administered
42	Utah Test of Language Development	Communication Research Associates, Inc., Box 11012, Salt Lake City, UT 84111	speaking; listening; general language ability	ages 2–14	individually administered
43	Vermont Basic Competency Program Speaking and Listening Assessments	Vermont Department of Education; Montpelier, VT 05602	speaking; listening	grades k–12	variety of simulation tasks and observations conducted in classrooms
44	Wallner Test of Listening Comprehension	N. K. Wallner, "The Development of a Listening Comprehension Test for Kindergarten and Beginning First Grade." *Educational and Psychological Measurement*, 1974, *34*, 391–396	listening	grades k–1	recorded instructions, listening passages, and responses; multiple choice format; administered to small groups.
45	Westside High School Minimum Competency Test	Westside Community Schools, Omaha, NE	speaking	grade 10	students present individual talks to group

Content domains

Since there exists no hegemony concerning the definition or nature of communication competence (Wiemann & Backlund, 1981), it is not surprising to find considerable diversity in associated measurement practices. A few assessment programs include measures of communication attitudes (26, 38). Others include tests of knowledge about communication, that is, indirect measures of communication competence (20, 26). A few procedures base evaluations on interactive communication in which speaking and listening alternate in a more or less naturalistic fashion (9, 10, 18, 37). Most tests, however, isolate speaking from listening.

The content domain of measures of speaking proficiency can be categorized according to (1) modes of discourse, (2) situations/audiences, and (3) evaluation criteria. At the elementary level, most speech assessment instruments elicit narrative (4, 38) or descriptive (9, 10, 18, 42) discourse. Story-telling (15, 17, 30) and description (16) are also featured prominently in communication tasks for non-native speakers. Older native speakers are most often called upon to deliver extended expository speeches (13, 20, 28, 45), although persuasion (13, 14, 22, 43) ritualized introductions (43), and other aspects of conversation (13, 22, 23, 43), appear occasionally among tests of speaking ability.

For all communication assessments, of course, the ultimate situation is evaluative and the ultimate audience is the examiner. Some procedures make no pretence otherwise, utilizing a single examiner-audience in interview (13, 22, 43) and even in extended speaking (13, 38) tasks. Experience suggests, however, that interviewers can exert overriding influence over the speech performance of examinees (Bazen, 1978; Mullen, 1978). Some procedures seek to mitigate the intrusiveness of the examiner-audience by placing students in simulated situations in which they role-play "life role" interactions (22, 14, 43). Dyadic referential accuracy exercises make use of peer audiences (9, 10, 18; See also Dickson & Patterson, 1981) as do some public speaking tasks in which speakers address their fellow students rather than examiners (28, 45).

Criteria for evaluating speaking performances also reflect divergent conceptions of the content domain. Becker (1962) found that typical analytic speech rating scales reflect three dimensions of judgment: content, delivery, and language. With the addition of organization, these dimensions account for most rating scales surveyed here. In particular, however, scales differ in their treatment of language. Some rating schemes award much credit to use of Standard American English patterns (13, 45). Other instruments, particularly those designed for non-native speakers, convey detailed information about the types of grammatical constructions mastered.

Indeed, certain speech assessment procedures appear to be more tests of productive language than of communication competence. Several tests, for example, require students to imitate words or sentences in isolation and then score performances solely for articulation accuracy or for evidence of first language interference (29, 39, 42). Even procedures which sample speech in communication situations, but which subject those speech samples to exclusively linguistic evaluation criteria (e.g. McCaleb, 1979; Mullen 1978) can not be construed as measures of communication proficiency. A few procedures sampled in this survey do suggest communication-oriented models of linguistic evaluation. Ratings may depend, for example, on the match between response and question type (23), or on the degree of elaboration as opposed to simple labeling (4, 17, 30).

The content domain of listening tests heavily emphasizes measures of literal comprehension (1, 3, 6, 8, 19, 20, 21, 24, 25, 26, 27, 28, 32, 33, 34, 35, 44). In recognition that literal comprehension scores may be confounded with recall skills, some listening instruments attempt to minimize this dependency by selecting brief passages coupled with few questions. Other skills frequently measured in listening tests include listening for directions (1, 11, 16, 42, 43), ascertaining the speaker's purpose (21, 25, 27), making inferences (3, 19, 20, 21, 25, 27, 28, 32), and summarizing (25, 43). Nonverbal decoding of paralinguistic cues is tested by several instruments (19, 31; also Davitz & Mattis, 1964; Smith-Elliot Listening Test, n.d.), as is decoding of visual cues (26,31; Smith-Elliot Listening Test, n.d.)

As with several of the speech examinations, many listening tests are more sensitive to narrow linguistic skills than to functional communication competence. Thus several measures are sensitive to receptive vocabulary (1, 2, 3, 6, 8, 15, 32, 39, 42), syntax (32) and phoneme recognition and discrimination (3, 11, 15, 29, 32, 33, 36, 39, 40). Listening tests for younger children, in particular, often fail to distinguish indicators of reading readiness from indicators of proficiency in receptive communication.

Response and scoring procedures

Multiple choice response formats are common among tests of listening ability (1, 2, 3, 4, 8, 19, 20, 21, 24, 27, 28, 32, 33, 34, 35, 40, 44). Multiple choice questions can measure literal comprehension most readily, but are also suitable for appraising higher order listening skills such as inference making and recognition of speaker's purpose. One difficulty inherent in most uses of multple choice questioning in listening tests is the necessity for students to read printed questions and response options. Procedures which provide tape recording of these materials, as well as tape recorded listening passages, minimize the confounding of reading and listening skills (21, 25, 26, 44). Pictorial, rather than verbal, response options (4, 33, 40) can also mitigate this confounding. In addition to multiple choice formats, some measures of listening ability feature items which demand behavioral responses to oral instructions (e.g., "Place a circle around the second largest square"; 1, 11, 16, 34, 42, 43).

Performance rating scales are the most common means for assessing speaking skill (4, 13, 14, 15, 16, 17, 20, 22, 18, 30, 38, 43, 45). Rubin (1981) summarizes pragmatic and psychometric factors pertaining to use of this technique in large scale testing programs. As an alternative to using performance rating scales, a number of procedures employ the incidence of particular discourse features as indicators of the quality of students' speech. Both Loban (1976) and McCaleb (1979), for example, suggest that syntactic complexity is an index of the quality of expression. Other discourse features utilized as quality indicators in measures of speaking ability are similarly linguistic in nature, e.g., total number of words, lexical diversity, articularion accuracy, and sentence expansion (4, 7, 16, 23, 37, 42). A few instruments use some combination of linguistic and whole-text (e.g., "Narrative goes beyond the information given in the pictoral stimulus") discourse features (17, 30). Such uses of discourse features as evaluative criteria of speaking ability are often proposed, however, without any evidence of concurrent validity which might demonstrate the strength of these measures as predictors of overall quality of expression. Indeed, syntactic complexity, in particular, has been shown to be unrelated to quality of expression in any direct way (Crowhurst, 1979).

Ideally, speaking proficiency should be measured by some indicator of communication effectiveness, of intended impact on an audience. Referential communication accuracy tasks (Dickson & Patterson, 1981) have been adapted in several instruments as measures of communication effectiveness (19, 10, 18). In these tasks speakers encode features of a target stimulus so that listeners can either discriminate that stimulus from others in an array, or else reproduce the stimulus. Speakers' accuracy (effectiveness) scores, however, may be contaminated by varying levels of listeners' decoding skills. Some referential accuracy scoring procedures, on the other hand, allow communication effectiveness to be ascertained by counting the number of criterial stimulus features, specificable *a priori,* which appear in speakers' descriptions (Piché, Rubin & Turner, 1980).

Administrative feasibility

Unlike other basic skills, communication is a social act. Tests of communication competence are therefore apt to be administratively more complex and expensive than many other large scale assessment procedures. Many measures of listening ability, however, are amenable to group administration (1, 2, 3, 5, 6, 8, 19, 21, 24, 25, 26, 27, 32, 35, 44), particularly when administration instructions and response options are tape recorded (21, 25, 26, 44). Those listening tests which are primarily diagnostic aids or which allow for a wide range of response modes, on the other hand, do require individual administration (11, 15, 16, 42).

Speech performance rating procedures naturally require individual administration. (No small group discussion tasks appeared in the instruments surveyed here. But see Follard & Robertson, 1976). While some procedures require multiple raters in order to enhance reliability (13, 14) others rely on a single rater or are ambiguous about rater requirements (22, 28, 38, 43, 45). A number of procedures alleviate allocation of resources by utilizing students' regular classroom teachers and classroom time for speech assessments (28, 38, 43, 45). One system employs a two-tier model (22; see also Carroll, 1980) in which classroom teachers first screen out those students who clearly achieve mastery as indicated by their typical communication behaviors. In the second stage, specially designated rater/administrators assess those students who did not display criterion performance in the first screening. Those speech performance assessment instruments which require that speech samples be transcribed into print (30, 37) place additional burdens on institutional resources, as do scoring systems which necessitate raters with special expertise in identifying linguistic structures (7, 16, 42).

Target populations and potential sources of test bias

The instruments reviewed here cover the entire K-12 age range, although the elementary grades receive particular emphasis, especially among commercially developed instruments. Several of the measures include alternate forms which can be administered in English or in Spanish (15, 16, 17, 30). Indeed, it appears that sophisticated advances in communication assessment have emerged from the field of second language testing (Carroll, 1980). Only a single instrument is specifically designated as appropriate for special education populations (11).

Stiggins (1981) discusses a number of sources of bias in communication testing. Instruments vary considerably in their efforts to minimize group bias effects. Some technical manuals document the work of minority group reviewers who examined items in order to eliminate potential bias (26). Other manuals tabulate normative data separately for black and white students (4). It should be noted, however, that

differences in central tendency are not, themselves, evidence of test bias. Rather, a test is biased if it over- or under-predicts scores on some independently administered criterion measure (Cleary, 1968). In the absence of criterion measures of communication quality it is difficult to ascertain test bias. The majority of instruments reviewed here, however, do not address the issue of potential group biases. Indeed, some scoring rubrics assign particular weight to standard English dialect patterns, a procedure which likely places nonstandard dialect speakers at a disadvantage.

Selected Research and Development Priorities

It is outside the purview of this report to recommend or approve particular assessment procedures. Indeed, the very construct of validity is situation specific; a given measure is valid only for specific purposes and specific populations (Cronbach, 1971). It seems advisable, therefore, that test adoption decisions be made on the basis of enlightened inquiry at local levels. However this review of available measurement procedures does serve to identify priorities for further research and development efforts in communication competence testing.

Delineating the content domain

It is not difficult to ascertain content validity in most tests of educational achievement. When learning objectives are identified, test constructors proceed to create a blueprint specifying the number and types of items addressed to each objective (Tinkelman. 1971). This mode of operation does not work well for speech communication testing for two reasons. First, in many cases oral communication instruction is not well established as a general education curriculum domain. Speaking and listening tests are therefore "proposing a wider range of curricular concerns in oracy than schools presently undertake" (Barnes, 1980, p. 125). Second, there is no consensually accepted conception of the communicaton competence construct, (Wiemann & Backlund, 1981) which could guide test construction in the absence of operationalized learning objectives. Indeed, the lack of conceptual clarity may be the greatest impediment facing communication assessment (Larson, 1978).

Various lists of communication competencies have proliferated (Allen & Brown, 1976; Basset, Whittington & Staton-Spicer, 1978; Edmonton Public School System, 1979). Testers who accepted the definitions of one or another of these documents were sometimes unable to devise suitable items for all of the components specified (McCaleb, 1979; Mead, 1977). Thus, it may not be feasible to test the entire domain of communication competence even when some conceptual scheme is adopted. In addition, it may not be desirable to test some components of communication competency that school officials may deem to be outside the proper purview of public education (e.g., self disclosure).

An especially troublesome issue pertaining to the content validity of oral communication measurement instruments concerns the role of language knowledge and general verbal ability. Effective communication requires the confluence of verbal, social, and logical abilities. It is at the same time a motor and perceptual skill, and is also influenced by attitudes. Measuring the simultaneous interaction of these subskills represents an especially elusive enterprise. Consequently assessment procedures may capture one or another of the more accessible components of communication skill, and most often this component is language. In the past, for example, commercially available listening tests have been criticized as little more than traditional reading tests presented orally, tapping general verbal ability more

than any unique aspect of the listening process (Kelly, 1965). Similar criticims may be leveled at any set of evaluation criteria which credit particular linguistic or stylistic features in an absolute fashion, criteria which may emphasize the socially prescriptive criterion of "correctness' but disregard the rhetorical criteria of intelligibility and appropriateness.

Accordingly, the following research and development priorities pertaining to content validity are proposed:

- Conduct comprehensive surveys of general education classroom learning objectives in oral communication.
- Devise a consensually acceptable conception of communication competence.
- Devise principles for sampling items from the content domain which are administratively feasible and appropriate within the realm of public education.
- Develop measures which distinguish between communication competence and general verbal ability or linguistic knowledge.

Establishing Criterion Measures of Communication Quality

Few of the instruments surveyed in this report have been subjected to studies of concurrent or predictive validity. That is, the degree to which test scores agree with some independent and well accepted criterion measure of communication quality remains largely unknown. The credibility of instruments which lack demonstrated criterion referenced validity is considerably diminished. Criterion referencing is particularly crucial for assessment tasks which are transparently contrived solely for the purpose of evaluation. A student, for example, may be asked to conduct a "conversation" with an interviewer/examiner. To what degree do the results of such tests reflect a student's competence in more naturalistic settings? Similarly, some assessment procedures require students to role-play familiar communication acts, but the relationship of students' performance in these simulated situations to their performance in real situations is obscure. Finally, it is not possible to ascertain group bias in tests without a criterion measure of communication quality against which test scores can be statistically regressed.

At present, however, there appear to be no well accepted criteria which can be used for validation purposes. Holistic teacher ratings of student's typical communication proficiency might prove suitable in this regard. Sociometric analyses using peer interaction data might also serve as criteria for establishing concurrent validity. Criteria for studies of predictive validity could include teacher or job performance ratings at some later point in time.

Accordingly, the following research and development priorities pertaining to criterion measures of communication quality are proposed:

- Establish criterion measures for ascertaining concurrent and predictive validity of assessment instruments.
- Explore the use of data gathered in naturalistic settings for purposes of criterion referencing.
- Determine the criterion referenced validity of contrived communication tasks.

Enhancing the reliability of measurements

Test scores are, of course, mere estimates of student ability. A host of extraneous factors may impinge on the accuracy of those estimates, that is, on reliability of measurement. Several of these factors are related to the characteristics of the

test-taker: time of day, mental and physical state, amount of prior "coaching." Few of the measures reviewed here report test/re-test reliability which might offer evidence concerning the impact of such characteristics on test peformance. Researchers in the evaluation of writing ability have advocated taking multiple writing samples in order to minimize the effects of temporary states on estimates of students' writing ability. (Braddock, Lloyd-Jones & Shoer, 1963; Deiderich, 1974). Because listening performance depends on attentional processes, and spoken messages include unintentional affect cues, it is likely that extraneous and temporary characteristics of test-takers affect reliability considerably (cf., Marine, 1965).

Other factors which affect error of measurement are inherent in the design of measurement procedures. Some speech proficiency examinations, for example, offer students a choice of topic under the assumption that students will avail themselves of the option which affords them the greatest comfort and fluency. Few such examinations, however, have ascertained whether each topic constitutes an "equivalent form" of the test. By the same token, interviewer-examiner idiosyncracies can gravely affect student performance in interview tasks (Bazen, 1978; Hitchman, 1966; Mullen 1978). When listening passages are read aloud on-site by test administrators, additional sources of extraneous score variance are intruded into listening tests.

Lack of agreement among judges is the bane of speech performance rating procedures. Considerable progress had been made in the field of composition evaluation in identifying sources of rater error and in devising training programs to enhance inter-rater reliability (Cooper & Odell, 1974; Diederich, 1974; Myers, 1980). Work in the area of speech evaluation, while promising, has not attained this level of sophistication (Rubin, 1981).

Most conceptions of communication competence hold that communication performance is situationally dependent, that competent communicators adapt their messages and their inferences to properties of the communicative context (Larson, et al, 1978; Allen & Brown, 1976). Therefore performance in one communication situation (dyadic, group, intimate, formal) may not reflect performance in another. Indeed, the most conceptually sound assessment strategy would sample student performance in a variety of contexts ("Criteria for Evaluating Instruments and Procedures for Assessing Speaking and Listening," 1979; Rubin, 1981). While some assessment programs do provide for cross-situational sampling, others limit tests to a single communication context in the interest of conserving resources. Little is known, however, about the extent to which performance elicited only a single communication context constitutes an accurate source of information about communication competence, in general.

Accordingly, the following research and development priorities concerning reliability of communication competence assessments are proposed:

- Ascertain test/re-test reliabilities of assessment instruments.
- Determine the degree of equivalence between various topics and tasks used in measures of speaking and listening.
- Devise procedures to minimize the impact of interactive test adminstrators and examiners.
- Investigate the role of contextual diversity, the number and types of communicaton situations, in ensuring adequate measurement of overall communication competence.

- Refine and disseminate methods for enhancing inter-rater reliability in speech performance ratings.

Identifying sources of test bias

As discussed previously, criterion measures of communication quality are necessary in order to demonstrate test bias, and such measures are currently in need of development. Nevertheless, culture bound evaluation materials will likely (but not necessarily) favor one cultural group over another. Such materials may include culture bound communication contexts (e.g., role-playing a business executive), test stimuli (e.g., "Point to the grandfather clock"), or evaluation criteria (e.g., standard English pronunciation, amount of eye-contact).

A more subtle source of bias against particular cultural groups may be inherent in the very notion of oral communication assessment. For some individuals, previous socialization may render communication testing an anomolous situation for which rules of appropriate behavior are lacking. Gay and Abrahams (1973), for example, contend that black youngsters generally construe and react to direct questioning by adults quite differently than do white middle class children. In a like manner, some Native American Indian children may have no basis for assimilating situations which call for individual noncooperative performances (Philips, 1970). Middle class children may understand that some questions are motivated by a genuine desire for information while other questions serve merely as a pretext to elicit displays of skill. Working class children, on the other hand, may be confused by "quasi-questions" whose actual purpose is to stimulate an evaluable performance (Bernstein, 1977). Moreover, some cultures place a premium on reserved speech (Hymes, 1974; Philipsen, 1975), in diametric opposition to the value system implicit in most communication assessment programs.

Increasingly large numbers of public school students are not native speakers of English. There is no reason why these students need be exempt from communication assessment, though they would, of course, be greatly disadvantaged if they were prohibited from using their native languages. Few instruments at present offer alternate forms appropriate for minority language groups however.

In additon to biases against particular linguistic and cultural groups, it is possible that communicaton assessment procedures may differentially treat certain exceptional populations. Surely provisions must be available for individuals who are handicapped by organic speech or hearing disorders. Policy decisions about the treatment of certain reticent individuals will also be needed. Shall students with personality traits like communication apprehension (McCroskey, 1977) be treated in a manner parallel to students with organic disorders? If not, and those students who experience communication apprehension *are* subjected to the same measurement procedures as the majority of students, then it would seem that public schools are thereby committed to "remediating" this condition as part of their responsibility to prepare students for competency examinations.

Accordingly, the following research and development priorities pertaining to bias in speaking and listening tests are proposed:

- Utilize criterion measures of communication quality to determine test bias.
- Identify culture bound communication contexts, evaluation criteria, and stimulus materials and estimate their contribution to measurement error.

- Subject to public policy analysis the conflict between norms inherent in communication assessment procedures and norms of communication held by cultural subgroups.
- Construct equivalent forms of communication tests for nonnative speakers.
- Construct guidelines for communication assessment of students with speech and hearing disorders.
- Clarify the status of dysfunctional personality traits like communication apprehension with respect to assessment requirements.

Innovating measurement techniques

Education agencies wishing to implement speaking and listening assessments have been unable to locate suitable instruments among extant procedures, and have therefore been compelled to engage in their own test construction efforts (e.g., Brown, et al., 1979; Plattor, et al, 1978). Although it may not be possible to anticipate or conform to objectives adopted in a variety of jurisdictions, the survey of assessment procedures reported here does suggest some gaps among the set of currently available models for testing communication competence.

Communication performance is highly dependent on situational features. To the extent that assessment situations deviate from naturalistic situations with genuinely communicative motivations, test scores will inadequately represent true communication skills. Yet few assessment schemes utilize data from naturalistic observation. Non-intrusive, naturalistic observation does, however, risk inconsistency in communication tasks, interactants, and perhaps rater expectations.

Conspicuous in their rarity are procedures which do not artificially isolate speaking from listening. Test of interaction skills, however, are liable to many of the same problems as naturalistic observation. That is, the control necessary for reliable measurement is difficult to achieve. Assessment of communication skill in the context of small group interaction may prove both psychometrically and administratively feasible (Barnes, 1980; Folland & Robertson, 1976, Becker, 1956). Referential communication accuracy tasks which permit free interaction between members of dyads (Dickson & Patterson, 1981) might likewise be adaptable for measuring communication skills in interaction.

The introduction of more diverse and realistic items in tests of listening skill represents another area for further development. A number of listening tests reviewed here already include items which call upon students to decode and make inferences about social interactions. The majority of listening assessment instruments, however, utilize monologic listening passages. No measures make provision for more active listening skills, such as formulating appropriate questions to pose to speakers. Little effort has been directed toward techniques which would measure ability at integrative decoding—using verbal, paralinguistic, and visual cues. Only one of the measures reviewed here, for example, presented videotaped listening passages.

Accordingly, the following research and development priorities pertaining to innovating measurement techniques are proposed.

- Develop non-intrusive assessment techniques utilizing naturalistic observation.
- Develop procedures for assessing interactive communication skills, including small group and dyadic interaction.
- Develop procedures which permit evaluation of active listening skills.

- Develop tests of integrative decoding ability which utilize verbal and nonverbal stimuli.

Appraising the utility of communication assessments.
It appears that communication assessments of one type or another can be implemented and it appears that certain educational benefits can accrue. It should not be assumed, however, that in every case it is desirable to assess communication competence. The utility of a test is a function of its costs and benefits (Cronbach & Glesser, 1965). Few inquiries have analysed the utility of measuring speaking and listening skills.

Costs bearing on the allocation of institutional resources are obvious: personnel costs for test administration and scoring, costs for consumable materials and for equipment, depletion of instructional time which is diverted to evaluation. Even these obvious costs, however, have rarely been quantified and reported. Other costs are less readily apparent. For example, it is possible that testing programs may result in deterioration of student attitudes, and that this deterioration may offset any positive learning outcomes that might otherwise be forthcoming.

The potential benefits of speaking and listening assessment programs are contingent on the uses to which test data are put. If data are used for certifying students' competence, then test utility will depend in large part on the quality of remedial instruction available for students who fail to demonstrate mastery on their first trial. Given the dubious psychometric adequacy of many of the procedures surveyed in this report, the costs of misclassifying students must also be taken into account. Test results may be used to evaluate programs, rather than individuals. In that case, the benefits of testing will depend upon the weight they are given in educational decision making.

In contrast to evaluating existing programs, programs of communication skills assessment may serve as a potent force in innovating new curricula. In arguing for oral language evaluation, Loban (1976, frontispiece) observed that "the language arts curriculum inevitably shrinks or expands to the boundaries of what is evaluated." Wilkinson (1968) forwards a similar claim about the "washback" to teaching of oral examinations, and Rubin (1980) projects that speaking and listening tests with "pedagogical validity" may facilitate sound curriculum and effective teaching practices. At present, such claims remain speculative, but do warrant expectations that communication competence assessment may, indeed, promote the quality of communication education in public schools.

Accordingly, the following research and development priorities pertaining to the utility of communication assessment are proposed:

- Estimate the institutional costs of various measurement procedures.
- Ascertain the effects of communication testing on student attitudes toward speech communication.
- Identify patterns of utilization of assessment data.
- Evaluate resources for providing remediative services to students found to be deficient in communication skills.
- Determine if assessment instruments are sensitive to instructional intervention.
- Investigate the effects of communication assessment programs on curricular innovation.

NOTES

[1]This article is an abridged version of the final report of the Committee on Assessment Instruments and Instrument Development (PreK-12) sponsored by the Speech Communication Association's Task Force on Assessment and Testing. This report, however, reflects only the views of the authors. The authors express appreciation to W. Patrick Dickson and to Janice Peterson, both of the University of Wisconsin, Madison, who assisted in reviewing the assessment instruments.

[2]No claim of comprehensiveness is made for the sample of assessment instruments. Nor does inclusion of any instrument in this compilation constitute an endorsement of any kind.

REFERENCES

Allen, R. R. & Brown K. L. (Eds.). *Developing communication competence in chidlren.* Skokie, IL: National Textbook, 1976.

Backlund, P. A national survey of elementary and secondary speaking and listening assessment. In R. Stiggins (Ed.), *Perspective on oral Communication Assessment in the 80's.* Portland, OR: Northwest Regional Educational Laboratory, 1981.

Barnes, D. Situated speech strategies: Aspects of the monitoring of oracy. *Educational Research,* 1980, *32,* 123–131.

Bassett, R. E., Whittington, N., & Staton-Spicer, A. The basics in speaking for high school graduates: What should be assessed? *Communication Education,* 1978, *27,* 293–303.

Bazen, D. The place of conversation tests in oral examinations. *English in Education,* 1978, *12,* 39–50.

Becker, S. L. Rating discussants. *Speech Teacher,* 1956, *5,* 60–65.

Becker, S. L. The rating of speeches: Scale independence. *Speech Monographs,* 1962, *29,* 38–44.

Bernstein, B. Foreword. In D. S. Adlam, *Codes in context.* London: Rutledge and Kegan Paul, 1977.

Braddock, R., Lloyd-Jones, R. & Schoer, L. *Research in Written composition.* Champaign, IL: National Council of Teachers of English, 1963.

Brown, K. L., Backlund, P., Gurry, J., & Jandt, F. *Assessment of basic speaking and listening skills* (2 vols.). Boston: Massachusetts Department of Education Bureau of Research and Assessment, 1979.

Buros, O. K. *The eighth mental measurement* yearbook (2 vols.). Highland Park, NJ: Gryphon, 1978.

Carroll, B. J. *Testing communicative performance:* An interim study. Oxford, U.K.: Pergamon Press, 1980.

Cleary, T. A. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement,* 1968, *5,* 115–124.

Cooper, C. R. & Odell, L. (Eds.). *Evaluating writing.* Urbana, IL: National Council of Teachers of English, 1977.

Criteria for evaluating instruments and procedures for assessing speaking and listening. *SPECTRA,* 1979, *15,* 5.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions.* (2nd ed.) Urbana, IL: University of Illinois Press, 1965.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational Measurement.* (2nd ed.) Washington, DC: American Council on Education, 1971.

Crowhurst, M. On the misinterpretation of syntactic complexity data. *English Education,* 1979, *11,* 91–97.

Davitz, J. R. & Mattis, S. The communication of emotional meaning of metaphor. In J. R. Davitz (Ed.), *The communication of emotional meaning.* New York: McGraw-Hill, 1964.

Dickson, W. P. and Patterson, J. H. Evaluating referential communication games for teaching speaking and listening skills. *Communication Education.* 1981, *30,* 11–21.

Diederich, P. B. *Measuring growth in English.* Urbana, IL: National Council of Teachers of English, 1974.

Edmonton Public School System. *Listening and speaking objectives: Divisions I-IV.* Edmonton, Alberta: Alberta Education, 1979.

Fagan, W., Cooper, C. & Jensen, J. Measures for research and evaluation in the English language arts. Urbana, IL: National Council of Teachers of English, 1975.

Follard, D. & Robertson, D. Towards objectivity in group oral testing. *English Language Teaching Journal,* 1976, *30,* 156–167.

Gay, G. and Abrahams, R. D. Does the pot melt, boil, or brew? Black children and white assessment procedures. *Journal of School Psychology,* 1973, *11,* 330–340.

Grommon, A. H. (Ed.). *Reviews of selected published tests in English.* Urbana, IL: National Council of Teachers of English, 1976.

Hitchman, P.J. CSE tests in spoken English, *Educational Research,* 1968, *10,* 218–225.

Hitchman, P. J. *Examining oral English in the schools,* London: Methuen, 1966.

Hymes, D. *Foundations in sociolinguistics: An ethnographic approach.* Philadelphia: University of Pennsylvania Press, 1974.

Johnson, O. G. *Tests and measurements in child development: Handbook II* (2 vols.). San Francisco: Jossey-Bass, 1976.

Kelly, C. M. An investigation of the construct validity of two commercially published listening tests. *Speech Monographs,* 1965, *32,* 139–143.

Lange, D. L. & Clifford, R. T. *Testing in Foreign languages, ESL, and bilingual* education, 1966–1979. (Language in Education: Theory and Practice, No. 24). Washington, DC: Center for Applied Linguistics/ERIC, 1980.

Larson, C., Backlund, P., Redmond, M., & Barbour, A. *Assessing functional communication.* Falls Church, VA: Speech Communication Association/ERIC, 1978.

Larson, C. E. Problems in Assessing functional communication. *Communication Education,* 1978, *27,* 304–309.

Loban, W. *Language development: Kindergarten through grade twelve.* (Research Report N. 18) Urbana, IL: National Council of Teachers of English, 1976.

Marine, D. R. An investigation of intra-speaker reliability. *Speech Teacher,* 1965, *14* 128–131.

McCaleb, J. Measuring oral communication. *English Education,* 1979, *11,* 41–47.

McCroskey, J. C. Oral communication apprehension: A summary of recent theory and research. *Human Communication Research,* 1977, *4,* 78–96.

Mead, N. A. *The development of an instrument for assessing functional communication competence of seventeen-year-olds.* Unpublished dissertation, University of Denver, 1977.

Mullen, K. A. Direct evaluation of rater and scale in oral interviews. *Language Learning,* 1978, *28,* 301–308.

Myers, M. *A procedure for writing assessment and holistic scoring.* Urbana, IL: National Council of Teachers of English, 1980.

Piché, G. L., Rubin, D. L., & Turner, L. J. Training for referential communication accuracy in writing. *Research in the Teaching of English,* 1980, *14,* 309–318.

Philipsen, G. Talking like a man in Teamsterville. *Quarterly Journal of Speech,* 1975, *61,* 13–22.

Philips, S. V. Acquisition of rules for appropriate speech usage. *Monograph Series on Language and Linguistics.* No. 23. Washington, DC: Georgetown University Press, 1970.

Plattor, E., Unruh, W. R., Muir, L., & Loose, K. D. *Test development for assessing achievement in listening and speaking.* Edmonton, Alberta: Alberta Education, 1978.

Richard, D. D. Communication competency assessment for non-native speakers of English. Paper presented at the Annual Meeting of the Southern Speech Communication Association, Austin, Texas, April, 1981.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. *Sensitivity to nonverbal communication: The PONS test.* Baltimore: Johns Hopkins University Press, 1979.

Rubin, D. L. Psychometric and pedagogical validity in large scale assessment of oral communication skills. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, 1980.

Rubin, D. L. Using performance rating scales in large scale assessments of speaking proficiency. In R. Stiggins (Ed.), *Perspectives on Oral Communication Assessment* in the 80's. Portland, OR: Northwest Regional Educational Laboratory, 1981.

Schools Council, *The Certificate of Secondary Education—trial examinations—oral English.* Examinations Bulletin No. 11. London: Her Majesty's Stationery office, 1966.

Smith-Elliot Listening Test, Needham, MA: Learning Dynamics, Inc., n.d.

Stiggins, R. J. Potential sources of bias in speaking and listening assessment. In R. J. Stiggins (Ed.), *Perspective on the assessment of speaking and listening skills for the 1980's.* Portland, OR: Northwest Regional Educational Laboratory, 1981.

Tinkelman, S. N. Planning the objective test. In R. L. Thorndike (Ed.). *Educational Measurement.* (2nd ed.) Washington, DC: American Council on Education, 1971.

Wade, B. Assessing oral abilities at 16+ *English in Education,* 1978 *12,* 52–61.

Wiemann, J. M. & Backlund, P. Current theory and research in communicative competence. *Review of Educational Research,* 1980, *50,* 185–199.

Wilkinson, A. The testing of oracy, In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach.* London: Oxford University Press, 1968.