DOCUMENT RESUME

ED 197 413 CS 503 246

AUTHOR	Mead, Nancy A.
TITLE	Assessing Speaking Skills: Issues of Feasibility, Reliability. Validity and Bias.
INSTITUTION	Education Commission of the States, Denver, Colo.
SPONS AGENCY	Massachusetts State Dept. of Education, Boston.
REPORT NO	ECS-00-SL-56
PUB DATE	Nov 80
CONTRACT	80-227
NOTE	37p.: Paper presented at the Annual Meeting of the Speech Communication Association (66th, New York, NY, November 13-16, 1980).
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Educational Assessment: *Evaluation Criteria; Evaluation Methods: *Measurement Techniques; Minimum Competencies: Secondary Education: *Speech Instruction: *Speech Skills: State Programs; Test Bias: Test Validity
IDENTIFIERS	Basic Skills Improvement Policy (Massachusetts): *Massachusetts

ABSTRACT

Focusing on the problems of assessing the speaking skills of secondary school students, this paper provides one example of how those problems were addressed in the Massachusetts speaking assessment. The paper identifies four requirements for measures of speaking skills: (1) feasibility, (2) reliability, (3) validity, and (4) freedom from bias. The discussion of these requirements is followed by a general discussion of issues related to developing measures of speaking skills. The paper concludes with a description of the Massachusetts speaking assessment and with data on the reliability, validity, and bias of the assessment instruments. (Author/FL)

ASSESSING SPEAKING SKILLS: ISSUES OF FEASIBILITY, RELIABILITY, VALIDITY AND BIAS

No. 08-SL-56

by
Nancy A. Mead

Education Commission of the States
1860 Lincoln Street
Denver, Colorado 80295

A paper presented at the Speech Communication Association Annual Convention, New York, New York, November 1980. The work upon which this paper is based was performed pursuant to Contract No. 80-227 of the Commonwealth of Massachusetts. It does not, however, necessarily reflect the views of that agency. The materials presented in the appendixes of this paper are copyrighted by the Massachusetts Department of Education. The materials are restricted for use in Massachusetts schools for the specific purpose of implementing the Massachusetts Basic Skills Improvement Policy and Regulations and may be duplicated by schools for that purpose. Circulation, distribution, duplication or use for other purposes requires written permission from the Massachusetts Department of Education.

INTRODUCTION

Current interest in assessing speaking skills of students can be traced to legislation and regulations of several federal and state educational agencies, policy and guidelines in certain school districts, and concern expressed by the general public. In response to this interest, a variety of test development efforts recently have been launched. For example, the American College Testing Program has incorporated oral communication skills into a newly developed basic competencies test for entering college students. The Alberta Education Ministry has developed speaking assessment instruments and conducted a provincewide assessment at selected elementary and secondary grades. The state of Vermont has developed guidelines for assessing speaking skills of elementary and secondary students which are being used by teachers statewide. The state of Massachusetts has developed two speaking assessment approaches for secondary level students which have been used in a statewide assessment and will be provided to school districts for their use in the future.

This paper focuses on the problems of assessing speaking skills of students in school and provides one example of how those problems were addressed in the Massachusetts speaking assessment. The paper identifies four requirements for measures of speaking skills. The requirements are followed by a general discussion of issues related to developing measures of speaking skills. The paper concludes with a description of the Massachusetts speaking assessment and data on the reliability, validity and bias of the assessment instruments.

REQUIREMENTS FOR MEASURES OF SPEAKING SKILLS

The assessment of speaking skills presents several unique measurement problems. First, the assessment must focus on behavior which is interactive and ephemeral. Communication competence is demonstrated in a person's interactions with other people. A good communicator must be skilled in both speaking and listening and must be able to adjust his or her messages based on feedback from other people. Speaking performance is a fleeting activity. Unless mechanical devices are used, there is no concrete record of what transpires. Secondly, the criteria for measuring communication competence are tied to cultural and situational variables. The judgment of competence is derived from the norms of a particula culture and may differ from culture to culture. For example speaking up in a group may be valued highly in one culture and may be considered inappropirate in another. Also, the expectations for communication behavior depend upon the context of the situation. Some communication behaviors are acceptable, even preferable, in some situations and are unacceptable in others. For example, using slang is often highly effective in a student's conversation with friends but it is often considered incorrect when it is used in the classroom.

Irrespective of the complexity of communication competence, an assessment of speaking skills must meet the traditional requirements for any measurement activity. These requirements include reliability and validity of the measurement instruments. Also measurement instruments must be free of bias. This issue is particularly important for measures of speaking skills because of the cultural and situational nature of communication competence. Finally, measurement instruments must be feasible. Again, this issue is particularly important for measures of speaking skills because the interactive and ephemeral nature of communication behavior poses some major feasibility problems.

A prerequisite for any instrument for assessing speaking skills of students is feasibility. In order for an assessment instrument to be practical for schools, it must be easy to use, within a reasonable amount of time, with a minimum of disruption. In many cases it is necessary that assessment procedures be designed for use by regular classroom personnel within the structure of normal school activity.

A second requirement is developing instruments which are reliable. There are many forms of reliability. However, the most important form of reliability for a speaking measure is inter-rater reliability. All raters must rate all students in the same way.

Validity is a third requirement for speaking assessment instruments. Validity may be measured in many ways. It is important that measures focus on communication skills and not related skills, such as sociability or general academic achievement. Also, different measures of communication competence should yield similar results.

A final requirement is developing measures which are free of bias. This means that the measure does not elicit different responses among subgroups which cannot be attributed to differences in skills. The types of speaking situations should not be tied to a particular cultural or sex perspective. Speaking tasks should not require special experience or knowledge that is not shared by all students. Ratings of students' performance should not be affected by the racial/ethnic background of the rater or of the student.

The four requirements discussed above provide a framework for exploring the issues related to developing measures of speaking skills. The issues include types of approaches for eliciting and rating speaking behaviors, rater objectivity and test administration.

ISSUES RELATED TO DEVELOPING MEASURES OF SPEAKING SKILLS

Naturalistic Versus Structured Approaches

Two competing approaches for assessing speaking skills are a naturalistic method or a structured method. A naturalistic approach suggests unobtrusive measurement of a student's spontaneous, interactive communication behavior as it takes place in a normal setting, such as a classroom. A structured approach suggests more obvious measurement of a student's communication behavior using more contrived speaking activities, which might or might not involve interaction with other students or the test administrator. One of the factors underlying a decision regarding these approaches is the desire to measure normal speaking performance or the desire to measure optimal speaking performance. A naturalistic approach measures a student's everyday speaking acitivity. A structured approach measures a student's best attempt to complete speaking tasks.

The naturalistic approach provides the most accurate measure of normal communication performance in realistic settings. However, it poses a variety of administration problems. First, there is no assurance that the activities of interest will occur during a given length of time. Furthermore, the observer must deal with a large amount of activity, only some of which is relevant for the assessment. Also, the observer must deal with a large number of students at one time, one or more of whom are the object of assessment.

The naturalistic approach also suffers from problems related to reliability and validity. The communication behavior of the student being observed may be influenced by the teacher or other students in the classroom (e.g., the student may dislike the teacher or may be dominated by a particularly vocal classmate.) It is also possible that the ratings of the observer may be influenced by factors other than communication ability (e.g., sociability or overall achievement level of the student).

The structured approach sets up specific communication tasks. However, these tasks may suffer from artificiality and may not assess normal communication performance. There are several ways to organize a structured approach and each method has associated advantages and disadvantages.

One structured approach is to set up a situation where a group of students are asked to interact with one another on a single task. This method reduces artificiality by providing interaction and a sense of audience. However, a problem in an interactive situation is that students may not have the opportunity to participate equally. Also, students may be affected by the racial/ethnic, sex or friendship characteristics of the group. To implement assessments in an interactive setting, the test administrator must rate more than one student simultaneously or rate the students after the fact (using audiotapes or videotapes).

Another structured approach is to set up a situation where several students are tested at the same time, with each student being given different but parallel communication tasks. Students take turns responding to the tasks. This provides a sense of audience and allows for efficient test administration. However, here again there is the possibility that the make-up of the group might affect a student's response. Also, it gives some students more rehearsal time or the possibility of modeling other students. To implement assessments using this method, the test administrator must keep track of several students responding to different tasks.

Lastly, it is possible to set up a structured situation where the test administrator gives a single student specific tasks. This method limits the student to a student-adult communication framework. This problem may be reduced by creating hypothetical situations. For example, it is possible to give a task to a student and ask the student to respond as if he or she were talking with another student. It is possible to increase the naturalism of the situation by having the test administrator introduce systematic probes or contingencies into the situation (e.g., simulate the questioning in an employment interview). Both of these approaches necessitate establishing an "as if" stance of acting. The one-on-one approach gives each student a chance to do his or her best without the influence of other students. However, for some students the student-adult situation may provoke more anxiety than the student-student situation. In implementing this approach, the test administrator may give the same tasks to all students. Also the test administrator may focus on one student at a time.

Holistic Versus Focused Ratings

Another distinction which may be made in assessment approaches is between holistic ratings which measure speaking performance globally and focused ratings which assess specific speaking skills. A decision related to which approach is taken in part depends on one's view of the nature of communication competence. If one sees communication competence as a unitary trait that is demonstrated in all communication situations, one would favor a holistic approach. If one sees communication competence as a set of subskills, one would favor a focused approach. It is widely accepted that communication is a highly complex and interactive skill. There are mounting arguments that communication skills can be categorized in terms of the nature of the situation and the function being served. This suggests that there might be unique skills related to various informal and formal situations and also unique skills related to various functions such as informing and persuading.

Another factor contributing to a decision about the type of ratings is the reliability of the approaches. A focused approach suggests rating scales which measure very specific definable behaviors, e.g., can be heard, cannot be heard. It is easy to establish inter-rater reliability with these types of scales. However, there is also a danger that these types of rating scales are trivial and do not represent the complexity of communication tasks. At the other extreme a holistic approach suggests rating scales which focus on very general behaviors which are difficult to define objectively, e.g., effective communication. Holistic ratings can be used reliably by raters. However, it is difficult to know exactly what criteria the raters are using to come up with their ratings and indeed different raters may be using different criteria and still be consistent with one another.

Rater Objectivity

An overriding concern, irrespective of the specific assessment approach, is the objectivity of the rater. The best way to deal with this concern is to develop rating scales and scoring guides which are explicit and leave no interpretation up to the rater. However, completely objective scales are difficult to create. Furthermore, it would be shortsighted to design easy-to-score but trivial rating schemes.

A critical aspect of rater objectivity is the relationship between the rater and the student. One aspect of this issue is whether or not the rater knows the student or is a stranger to the student. A person who knows the student brings a particular sensitivity and understanding to the testing situation. However, it is possible that such a person might also bring preconceived notions to the testing situation. Using a stranger lessens this possibility. It is unclear whether a student will be more at ease with a familiar person or a stranger. Some might be more comfortable with their teacher, others more comfortable with a stranger.

Another aspect of the relationship between the rater and the student which may affect rater objectivity is the respective racial/ethnic backgrounds of the rater and the student. It is possible that raters tend to rate students of the same racial/ethnic backgrounds higher than students of different racial/ethnic backgrounds. It is usually impossible to match rater and student on all possible confounding factors such as race/ethnic, sex or socioeconomic background.

A third concern related to objectivity of the rater is the presence of a halo effect. In any type of rating approach (whether the rating is done on the spot or later) there is a danger that the rater's experience with previous students may affect how he or she rates subsequent students. For example, a mediocre response after many poor responses might be rated higher than a mediocre response after an exemplary response.

A related problem is fatigue and boredom. These problems are more readily dealt with in a situation where ratings are done after the fact, where regular breaks and consistency and calibration checks may be scheduled. In a situation where ratings are done on the spot, the best solution involves random ordering of test takers, reasonable time schedules and regular reviews of training material.

Test Administration

A final set of issues which concerns all structured assessment approaches deals with various aspects of test administration -- where the student will be tested, for how long and when student responses will be rated. From a management standpoint it is desirable to test students within their own classroom. However, this poses problems of disruptions by students not involved in the testing and a lack of privacy.

Likewise, the length of testing is a major management issue. All structured assessment approaches require setting aside time specifically to test an individual or small groups of students. Even with a very short test, this represents a major level of effort. Fortunately, even a relatively short assessment (e.g., five minutes), yields considerable information about speaking ability.

A final concern is with the rating approach. Student responses may be rated on the spot or they may audiotaped or videotaped for later scoring. The advantages of on-the-spot ratings are mostly practical ones. If test administrators can be trained to be reliable scorers, it is most time efficient to have them do the ratings at the same time as the administrations. Videotaping provides a good record of communication responses. However, it is quite expensive. Audiotaping provides a less complete record of communication responses; it does not provide visual information. However, it is much cheaper and easier to use.

A DESCRIPTION OF THE MASSACHUSETTS SPEAKING ASSESSMENT

The Commonwealth of Massachusetts recently adopted the Massachusetts Basic Skills Improvement Policy which among other things called for the development of assessment instruments and the conduct of statewide assessmonts in the basic skills. Basic skills were defined as mathematics, reading, writing, listening and speaking. The Education Commission of the States (ECS), under contract with the Massachusetts Department of Education, was responsible for developing instruments and conducting a statewide assessment of listening and speaking during the 1979-80 school year. The following describes the instruments which were developed for the speaking assessment and presents evidence related to, the reliability, validity and bias of the instruments. The speaking assessment is unique in that it uses a two-staged approach which is feasible for large-scale assessment in the schools within limitations of normal personnel and time resources.

Objectives and Assessment Design

The speaking assessment was designed to measure the 14 speaking objectives identified by the Massachusetts Basic Skills Improvement Policy. Exhibit 1 lists the speaking objectives and indicates how the speaking rating scales and tasks developed for the assessment are related to the objectives. Some of the speaking objectives deal with general speaking skills that apply to all speaking situations -- for example, Objective A-1, Use Words and Phrases Appropriate to the Situation. Other objectives deal with specific speaking situations -- for example, Objective C-1, Use Survival Words to Cope with Emergency Situations.

A set of speaking rating scales was developed for the speaking assessment. These scales were used in two ways:

- For classroom observation by teachers
- For one-on-one assessment by trained raters

EXHIBIT 1. Relationship of Tasks and Ratings to Speaking Objectives

Objectives	Tasks	Ratings
A. Basic Oral Communication Skills		
1. Use words and phrases appropriate to the situation		Content
2. Speak loudly enough to be heard by a listener or group of listeners		Delivery
3. Speak at a rate the listener can understand		Delivery
4. Say words distinctly		Delivery
B. Planning, Developing, and Stating Spoken Messages		
1. Use words in an order that clearly expresses the thought		Organization
2. Organize main ideas for presentation		Organization
3. State main ideas clearly		Content
4. Support main ideas with important details		Content
5. Demonstrate knowledge of standard English usage		Language
C. Common Uses of Spoken Messages		
1. Use survival words to cope with emergency situations	Respond to an Emergency	
2. Speak so listener understands purpose		Content
3. Ask for and give straightforward information	Explain a Sequence	
4. Describe objects, events and experiences	Describe an Object, Event or Experience	
5. Question others' viewpoints	Persuade Someone to Do Something	

The following sections describe the speaking rating scales, the teacher observation approach, the one-on-one approach and how the approaches are combined into a two-staged assessment.

Speaking Rating Scales

A set of four scales was developed to measure the objectives that deal with general speaking skills. Each scale measures one dimension of speaking skills. The dimensions include:

- Delivery
- Organization
- Content
- Language

The delivery dimension focuses on Objectives A-2, A-3 and A-4 and is concerned with how well a student transmits messages. It measures how well a student uses appropriate volume, rate and articulation while speaking. The organization dimension focuses on Objectives B-1 and B-2 and is concerned with how well a student structures messages. It measures how well a student expresses the sequence or relationships of ideas. The content dimension focuses on Objectives A-1, B-3, B-4 and C-2 and is concerned with how well a student provides an adequate amount of relevant information to meet the requirements of various speaking tasks. In addition, it measures how well a student adapts the content of messages to specific listeners and situations. The language rating focuses on Objective B-5 and is concerned with how well a student uses appropriate grammar and vocabulary while speaking.

Within each speaking dimension, performance is rated using a 4-point scale:

1 = Inadequate
2 = Minimal
3 = Adequate
4 = Superior

The dimensions and the levels of the rating scales are explained in the "Speaking Assessment Ratings Guide," which may be found in Appendix A.

The Teacher Observation Approach

The teacher observation approach is a general measure of a student's speaking performance. In this approach, two teachers who had the same student currently enrolled in class independently rated the student's general speaking performance in class using the speaking rating scales. Usually one of the teachers was an

English teacher and the other from another subject area where students participate in a fair amount of classroom discussion and interaction, such as history, government or science.

Teachers were asked to read the "Speaking Assessment Ratings Guide" and then to complete the ratings. Teachers based their ratings of a student on their observation of the student's performance in normal classroom activities, such as asking questions, responding to questions, explaining how to do something, giving a report to the class or talking with other students in discussion groups. They considered the student's average performance since the beginning of the semester.

The One-on-One Approach

The one-on-one approach is a focused measure of a student's speaking performance. In this approach, a trained rater rated the student's performance using the same speaking rating scales used by teachers for the teacher observation approach. However, instead of basing ratings on classroom observations, the rater gave a student specific tasks and rated the student's performance on these tasks.

Each rater was provided with one day of training in the one-on-one approach. The rater assessed each student individually. The rater gave the student several speaking tasks and rated the student's performance on each task along the four dimensions. Thus, for each task, the rater gave the student a rating from 1 to 4 for delivery, organization, content and language. The rating was assigned immediately after the student's response.

The speaking tasks used in the one-on-one approach reflect the objectives that deal with specific speaking situations. The tasks include:

- A description
- An emergency
- A sequence
- A persuasion task

The description task focuses on Objective C-4 and is concerned with how well a student can describe an object, event or experience so that another person would know something about the topic. The emergency task focuses on Objective C-1 and is concerned with how well a student can provide the necessary information in an emergency so that another person could send help. The sequence task focuses on Objective C-3 and is concerned with how well a student can explain a sequence of steps so that another person would understand the sequence. The persuasion task focuses on Objective C-5 and is concerned with how well a student can present effective arguments so that another person would be persuaded by the student's point of view.

The focus of all of the speaking tasks was on the student's effectiveness in transmitting the message not on the specific content of the messages. Tasks were developed which were familiar to all students and did not require any special knowledge or experience. The tasks were field tested with 9th graders and 12th graders from inner city, suburban and rural schools to assure that they would be relevant for a wide variety of students of various ages. Based on field test results, four tasks, one of each type, were selected for the final version of the speaking test. These tasks were judged to be substantially free of sex, racial/ethnic, geographic and socioeconomic bias by review committees established by the Department of Education. The text of the tasks is provided in Appendix B.

The Two-Staged Approach

The rationale for developing a combined approach for assessing speaking is based on the need for measures that are reliable, valid and free of bias. The teacher observation approach provides a general measure of student performance. It assesses all types of speaking tasks as they occur in a natural situation. However, sometimes a general measure such as this one allows for other factors, such as academic achievement or sociability, to enter into the ratings. The one-on-one approach provides a focused measure of student performance. It assesses only a few types of speaking tasks in a contrived setting. However, it focuses entirely on speaking variables and uses standardized procedures. The two approaches complement one another and taken together they guard against the many problems of reliability, validity and bias that are inherent in speaking measures.

Another reason for the combined approach is the need for measures that are feasible for large scale use by school districts in the future. The intention is to use the teacher observation approach as a screening measure and to use the one-on-one approach as a back-up measure in cases where a student's level of ability is in question. School districts would use the teacher observation measure to assess all students in a grade. They would use the one-on-one measure to assess students whose level of ability is in question. For example, the one-on-one measure might be used when two teachers do not agree in their observation ratings of a student.

Before final plans are made for implementing the speaking measures in school districts, it is necessary to demonstrate that the two approaches are reliable and valid. Initial tests of the reliability and validity of the measures were conducted during the statewide assessment and are presented in this paper. The review committees felt that initial results were encouraging. However, further study will be undertaken before final recommendations are made regarding how measures should be used school districts.

RELIABILITY OF THE SPEAKING RATINGS

Reliability refers to the degree of accuracy of the measurement process. There are many ways to examine the reliability of a measure. The most important aspect of reliability for the speaking assessment is inter-rater reliability. It is essential that teachers and raters rate the same student in the same way.

Reliability of the Teacher Observation Approach

Using the assessment results, it is possible to estimate the reliability of the teacher observation ratings. Table 1 indicates the percentage of agreement between teachers rating the same student for each speaking dimension. For example, for delivery 53.4% of the ratings are identical. Approximately 95% of the ratings are either identical or adjacent (within one point of one another). The level of agreement is consistent across all four dimensions.

The responses of teachers were also examined to see if there were systematic differences in the ratings of different types of teachers. For the most part one of the teachers rating a student (Teacher 1) was an English teacher and the other (Teacher 2) was a teacher from a subject area other than English. In effect, each teacher assigned an observation score to a student. The score is the sum of the ratings across all four dimensions and may range from 4 (a student receives 1 for all four dimensions) to 16 (a student receives 4 for all four dimensions). Table 2 indicates the mean teacher observation scores assigned by Teacher 1 and Teacher 2. Although the difference between the two means, -.023, is statistically significant at the .05 level, it is not practically significant. Furthermore, the responses of English and speech teachers to the Teacher Questionnaire were very similar to the responses of other teachers, indicating that all teachers reacted similarly to the teacher observation approach.

The review committees felt that the assessment results indicate an adequate level of agreement between teachers and

TABLE 1. Percentage of Identical Adjacent and Other Discrepant Teacher Observation Ratings for Each Dimension

	Percentage			
	Delivery	Organization	Content	Language
Identical ratings	53.4	55.8	55.6	56.8
Adjacent ratings (within one point)	41.9	41.9	39.8	41.0
Other discrepant ratings (more than one point apart)	4.8	2.4	4.8	2.4

Number of cases = 502

TABLE 2. Means and Mean Difference between Teacher 1 Observation Score and Teacher 2 Observation Score (Teacher 1 Score Minus Teacher 2 Score)

	Teacher 1 Score		Teacher 2 Score
Mean	11.570		11.773
Standard Deviation	2.130		2.137
Standard Error	0.095		0.095
Mean Difference		-0.023*	
Standard Deviation		2.291	
Standard Error		0.102	
Number of Cases		502	

suggested that teacher observation scores of two teachers be added together to obtain a total teacher observation score for a student. The total teacher observation score has a range from 8 to 32.

Reliability of the One-on-One Approach

Using the results from training and the assessment, it was possible to estimate the reliability of the one-on-one ratings.[1] Six individuals were trained as one-on-one assessment raters. All six had public school experience. The group consisted of three males and three females; it included three whites, two blacks and one Hispano. The group was selected to allow study of the interaction of sex and racial/ethnic background of raters and students in the one-on-one ratings. The six raters received one day of training in the one-on-one approach. They studied the same "Speaking Ratings Guide" used by teachers in the teacher observation approach. In addition, raters listened to tape recordings of students responding to the speaking tasks, attempted ratings and discussed results. At the end of the training session they independently rated tape recordings of five students, each student responding to four tasks. At this time the raters assigned identical ratings 85% of the time.

After the assessment, tape recordings of 10% of the student responses were rescored by project staff who also were reliable raters. Raters assigned identical ratings 75% of the time. The lower level of consistency on the rescoring may be due to poor quality of some tape recordings. Also, the experience of raters suggests that ratings made on the spot may be somewhat different than ratings made from tape recordings. There may be a tendency to be more attracted to the enthusiasm and presence of students when rating on the spot than when rating tape recordings.

The ratings of individual raters were examined to see if there were any systematic differences among raters. It is impossible to determine precisely whether or not raters were rating at the same level because they were assigned to schools based on geographical convenience, not randomly. However, average ratings tended to vary somewhat from rater to rater, ranging from 2.77 to 2.95. This suggests that there may be some tendency for individuals to be low raters or high raters.

[1] In addition, four individuals from the Department of Education and project staff were trained as alternates.

The data suggest that it is possible to establish a high level of reliability among raters after a relatively short training period and that raters maintain a fairly high level of reliability during an assessment period. Reliability might be improved if all student responses were tape recorded and scored at a later time during which maintenance of reliability could be monitored on an ongoing basis.

VALIDITY OF THE SPEAKING RATINGS

The validity of measurement instruments is concerned with the extent to which instruments actually measure the skills they are intended to measure. Validity may be determined by several methods. One important test of validity for performance measures is content validity. Content validity indicates the degree to which the content of a test represents the domain of skills it intends to measure. Content validity is usually determined through expert judgment.

In order to measure the content validity of the speaking measures, a panel of communication experts was assembled to review the measures. Prior to meeting, individuals received the speaking objectives and a general description of the measures. They were asked to categorize the speaking tasks and speaking dimensions according to the speaking objectives. The results of the categorizing were then discussed at a meeting of the panel. In general the panel agreed almost unanimously with respect to the categorizing of tasks according to the objectives and agreed most of the time with respect to the categorizing of dimensions according tc the objectives. The overall judgment of the group was that the measures generally reflected the objectives, although they felt that the measures did not directly assess asking for information or questioning another person's viewpoint. Based on the reviewers comments, adjustments were made in tasks and dimension so that they more nearly reflected the objectives.

Another test of validity which is appropriate for the present study is concurrent validity. Concurrent validity indicates the degree to which individuals respond to different measures of the same skill in the same way. The fact that two approaches were used for the speaking assessment made it possible to assess the concurrent validity of the approaches. In order to show the concurrent validity of the two speaking approaches, it is necessary to demonstrate that teachers and trained raters rate the same student in approximately the same way.

In order to determine the concurrent validity of the two approaches, two scores were compared -- the total teacher observation score and the adjusted total one-on-one score. The total teacher observation score is obtained by adding together the four ratings of the first teacher and the four ratings of the second teacher. This score ranges from 8 to 32. The total one-on-one score is obtained by adding together the four ratings of the trained rater on each of the four tasks on the speaking test. This score ranges from 16 to 64. The total one-on-one score is then divided by 2 to obtain the adjusted total one-on-one score, which has a range identical to the total teacher observation score.

The relationship of the total teacher observation score and the adjusted total one-on-one score is presented in Table 3. The table shows the percentage of scores which were identical or within 1 to 11 points of one another. The results indicate that 80.5% of the scores were within 4 points of one another and that 98.2% of the scores were within 8 points of one another. This suggests that most of the time individual ratings were, on the average, within 1/2 point of one another and that virtually all of the time individual ratings were, on the average, within 1 point of one another.

Additional evidence of the concurrent validity of the two speaking approaches is the fact that they yield the same overall mean. Table 4 indicates that there is no significant difference in the mean for the total teacher observation score and the mean for the adjusted total one-on-one score for the same students.

The review committees felt that the assessment results confirmed the validity of the two-staged speaking approach. However, they felt that some evidence relating to possible bias in the measures warranted further study. The aspects of bias are discussed in the next section.

BIAS IN THE SPEAKING RATINGS

Speaking assessments of all types are particularly vulnerable to bias. This is because speaking competence is to some extent determined by cultural and situational norms. In other words, what is considered competent performance differs from culture to culture and from situation to situation. As indicated previously, the teacher observation ratings are particularly susceptible to bias because they are a general measure of speaking performance. It is possible that teachers might take into account factors other than speaking skill in the ratings of a student's performance such as academic achievement or sociability. Also teachers might make judgments of a student's performance based on their personal

TABLE 3. Absolute Difference Between Total Teacher Observation Score and Adjusted Total One-on-One Score

Absolute Difference	Frequency	Percentage	Cumulative Percentage
0	57	11.7	11.7
1	125	25.6	37.3
2	96	19.7	57.0
3	60	12.3	69.3
4	55	11.3	80.5
5	39	8.0	88.5
6	21	4.3	92.8
7	17	3.5	96.3
8	9	1.8	98.2
9	6	1.2	99.4
10	2	0.4	99.8
11	1	0.2	100.0
Total	488	100.0	

TABLE 4. Means and Mean Difference between Total Teacher Observation Score and Adjusted Total One-on-One Score (Total Teacher Observation Score Minus Adjusted Total One-on-One Score)

	Total Teacher Observation Score		Adjusted Total One-on-One Score
Mean	23.299		23.357
Standard Deviation	3.576		2.156
Standard Error	.162		.098
Mean Difference		-0.057	
Standard Deviation		3.473	
Standard Error		0.157	
Number of Cases		488	

preference for a particular communication style, even though other styles might be equally effective.

The results of the speaking assessment provided some evidence of possible test bias. Initial analyses of the teacher observation and the one-on-one results indicated that more statistically significant group differences were identified in the teacher observation ratings than in the one-on-one ratings. Upon further examination, it was clear that most of the group differences identified by both approaches were in the same direction but that the differences in the teacher observation ratings were just enough larger that they fell into the category of statistically significant differences. For example, a group difference in the teacher observation ratings would be just over 2 standard errors and the same group difference in the one-on-one ratings would be about 1.5 standard errors.

However, there still were a few anomalies in the speaking results. For the teacher observation approach, the mean rating for whites was slightly above the statewide mean (a significant difference) and the mean rating for blacks was below the statewide mean (not a significant difference). For the one-on-one approach, the mean rating for whites was slightly below the statewide mean (not a significant difference) and the mean rating for blacks was slightly above the statewide mean (not a significant difference). It was felt that these anomalies warranted additional analysis.

With the existing data it is not possible to conclusively state whether or not either of the speaking approaches exhibits racial/ethnic bias. Data are available on the sex and racial/ethnic background of the students participating in the assessment. However, no data are available on the sex or racial/ethnic background of teachers who provided observation ratings. Some relevant data are available regarding the trained raters.

In exploring the data on the teacher observation measure, consideration was given to the reliability of the two ratings. It was hypothesized that ratings which were more reliable might also be less biased. High reliability was defined as instances where a student received ratings from two teachers and the ratings on each dimension were within one point of one another. Several analyses were conducted to explore the differences in results when all teacher observation ratings are included in the analysis and when only highly reliable ratings are included. Table 5 indicates the results of the teacher observation measure for all students and by racial/ethnic groups when all ratings are analyzed and when only highly reliable ratings are analyzed. The results indicate that group differences for highly reliable ratings are smaller in magnitude but in the same direction as the group differences for

TABLE 3-5. Mean Ratings for All Teacher Observation Ratings and for Teacher Observation Ratings All Within One Point for All Students and by Racial/Ethnic Groups

	All	White	Black	Hispanic	Other
All Teacher Observation Ratings					
Mean Rating	2.920	2.933	2.831	2.605	2.906
Standard Error	0.028	0.030	0.060	0.119	0.149
Mean Difference	--	0.014*	-0.089	-0.315*	-0.014
Standard Error	--	0.007	0.060	0.117	0.144
Number of Cases	560	487	31	16	9
Teacher Observation Ratings All Within One Point					
Mean Rating	2.945	2.946	2.941	2.756	3.041
Standard Error	0.031	0.033	0.073	0.095	0.172
Mean Difference	--	0.001	-0.004	-0.189	0.096
Standard Error	--	0.005	0.076	0.095	0.171
Number of Cases	452	401	24	9	6

all the ratings. No statistically significant group differences are found when the highly reliable ratings were analyzed.

Further analysis of the teacher observation approach included an examination of the differences in the total teacher observation scores and the adjusted total one-on-one scores when all teacher observation ratings are used in the analysis and when only highly reliable ratings are used. Table 6 indicates the mean difference scores for all students and for racial/ethnic groups for the two types of analyses. The optimum difference between the two scores is zero. This indicates that students received the same value for the total teacher observation score and the adjusted total one-on-one score. The results indicate that all mean differences are close to zero for all students and for all racial/ethnic groups of students for both types of analyses. None of the differences is statistically significant. Thus, these data indicate no systematic differences in scores obtained from the teacher observation ratings and the one-on-one ratings and the consistency is sustained when racial/ethnic groups are examined and when all teacher observation ratings and highly reliable teacher observation ratings are examined separately.

Analysis of the one-on-one results were conducted to look at the interaction between the racial/ethnic background of the rater and the racial/ethnic background of the student. Trained raters were selected specifically to reflect a variety in sex and racial/ethnic backgrounds. However, there were only six raters -- three males and three females; three whites, two blacks and one Hispano. They werc assigned to schools based on geographical convenience, not randomly. The number of minority students which each individual rater rated was small. Finally, an anomaly of the situation was that one black rater tended to be a very low rater and the one Hispanic rater tended to be a moderately low racer when compared with the other raters. All these factors confound exploratory analysis of the data. Table 7 indicates the mean ratings of students by the racial/ethnic background of the raters and the students. The number of students included in each mean rating is presented in parentheses after each mean rating. There is some evidence that white raters rated black students slightly lower and Hispanic students somewhat higher than white students. Black raters rated black students quite a bit higher and Hispanic students somewhat lower than white students. The Hispanic rater rated black students slightly lower and Hispanic students quite a bit lower than white students. It should be strongly emphasized that these data are subject to many contaminating factors, as indicated above.

The results of the exploratory analyses of the speaking measures with respect to possible racial/ethnic bias are inconclusive but they generally support the fact that both measures seem to be measuring the same skills and that the

TABLE 6. Mean Differences Between Total Teacher Observation Score and Adjusted Total One-on-One Score by Teacher Observation Ratings Used in the Analysis and by Racial/Ethnic Background of Students (Total Teacher Observation Score Minus Adjusted Total One-on-One Score)

	All	White	Black	Hispanic	Other
All Teacher Observation Ratings					
Mean Difference	-.057	.023	- .385	-1.071	- .444
Standard Deviation	3.473	3.483	3.910	3.025	3.678
Standard Error	.157	.169	.767	.808	1.226
Number of Cases	488	426	26	14	9
Teacher Observation Ratings All Within One Point					
Mean Difference	.096	.159	- .087	-1.000	.333
Standard Deviation	3.475	3.487	3.930	2.739	3.777
Standard Error	.166	.177	.820	.913	1.542
Number of Cases	439	389	23	9	6

TABLE 7 Mean One-on-One Ratings for All Students and Students by Racial/Ethnic Groups and for All Raters and Raters by Racial/Ethnic Groups

Raters	Students				
	All	White	Black	Hispanic	Other
All	2.89	2.89	2.92	2.60	2.92
	(691)	(600)	(34)	(19)	(10)
White	2.94	2.94	2.93	3.05	2.87
	(319)	(277)	(18)	(5)	(5)
Black	2.84	2.84	2.94	2.50	2.97
	(259)	(232)	(10)	(9)	(5)
Hispanic	2.85	2.85	2.83	2.71	--
	(113)	(91)	(6)	(5)	--

† Number of cases are indicated in parenthesis

measures are consistent across racial/ethnic groups. Some concern with the reliability of the teacher observation ratings has led to the recommendation that only instances where a student receives ratings from two teachers and that the ratings on each dimension are within one point of one another be considered usable ratings. Analyses in the summary and technical reports of the study only include teacher observation ratings which meet these criteria.

SUMMARY

Developing instruments for assessing speaking skills poses many challenges. Not the least of these challenges is creating strategies which are feasible for fairly large-scale use in schools. Added to this are the traditional measurement requirements that instruments be reliable, valid and free of bias. This paper discussed some of the issues related to assessing speaking skills, including a naturalistic versus a structured approach, holistic versus focused ratings, rater objectivity, and test administration. The paper also described one assessment approach developed for the state of Massachusetts. Any assessment of speaking skills is subject to a variety of measurement problems, many of which are inherent in the nature of communication competence. However, data suggest that the Massachusetts approach provides a feasible way of assessing speaking skills while still maintaining measurement requirements of reliability, validity and freedom from bias.

APPENDIX A

MASSACHUSETTS DEPARTMENT OF EDUCATION ASSESSMENT OF BASIC SKILLS

SPEAKING ASSESSMENT RATINGS GUIDE

OVERVIEW

There are numerous kinds of speaking tasks that students must perform in everyday life, both in school and out of school. The Massachusetts Basic Skills Improvement Policy has focused on some of these tasks, including describing objects, events and experiences, explaining the steps in a sequence, providing information in an emergency and persuading someone.

In order to accomplish a speaking task, the speaker must formulate and transmit a message to a listener. This process involves deciding what needs to be said, organizing the message, adapting the message to the listener and situation, choosing language to convey the message and finally delivering the message. The effectiveness of the speaker may be rated in terms of how well the speaker meets the requirements of the task.

The Massachusetts test of basic skills in speaking separates speaking skills into four dimensions:

Delivery
Organization
Content
Language

Delivery is concerned with the transmission of the message, i.e., volume, rate and articulation. Organization is concerned with how the content of the message is sequenced and how the ideas are related to one another. Content is concerned with the amount and relevance of information in the message, and how the content is adapted to the listener and situation. Language is concerned with the grammar and words which are used to convey the message.

Each of the four dimensions is rated on a four point scale: 1 is the lowest rating and 4 is the highest rating. A general set of principles underlies the rating scale for all four components. Ratings of 1 reflect speaking skills which are inadequate in meeting the requirements of the task. Ratings of 2 reflect speaking skills which are minimal in meeting the requirements of the task. Ratings of 3 reflect speaking skills which are adequate in meeting the requirements of the task. Ratings of 4 are superior in meeting the requirements of the task.

Individuals who act as raters for the speaking assessment need to take the role of a naive, objective listener. The rater must be naive so that the rater can base his or her rating on exactly what the speaker says. The rater must be careful not to let his or her own knowledge and experience influence the rating. The rater must face each speaker as if it were a new experience. The rater must also be objective so that he or she does not let a particular set of norms of social acceptability influence the rating. The rater must evaluate the speaker in terms of how well the speaker meets the requirements of the speaking task, irrespective of the particular communication style the speaker uses.

DELIVERY

The delivery rating focuses on the transmission of the message. It is concerned with volume, rate and articulation. Articulation refers to pronunciation and enunciation. Some examples of poor articulation include mumbling, slurring words, stammering, stuttering and exhibiting disfluencies such as ahs, uhms or "you knows."

1 = **The delivery is inadequate in meeting the requirements of the task.**

e.g., The volume is so low that you cannot understand most of the message.
The rate is so fast that you cannot understand most of the message.
The pronunciation and enunciation are so unclear that you cannot understand most of the message.

2 = **The delivery is minimal in meeting the requirements of the task.**

e.g., The volume is too low or too loud.
The rate is too fast or too slow. Pauses are too long or at inappropriate spots.
The pronunciation and enunciation are unclear. The speaker exhibits many disfluencies such as ahs, uhms or "you knows."
You are distracted by problems in the delivery of the message.
You have difficulty understanding the words in the message. You have to work to understand the words.

3 = **The delivery is adequate in meeting the requirements of the task.**

e.g., The volume is not too low or too loud.
The rate is not too fast or too slow. Pauses are not too long or at inappropriate spots.
The pronunciation and enunciation are clear. The speaker exhibits few disfluencies, such as ahs, uhms and "you knows."
You are not distracted by problems in the delivery of the message.
You do not have difficulty understanding the words in the message.

4 = **The delivery is superior in meeting the requirements of the task.**

e.g., The speaker uses delivery to emphasize and enhance the meaning of the message. The speaker delivers the message in a lively, enthusiastic fashion.
The volume varies to add emphasis and interest.
Rate varies and pauses are used to add emphasis and interest.
Pronunciation and enunciation are very clear. The speaker exhibits very few disfluencies such as ahs, uhms or "you knows."

NOTE: In articulation you may be concerned with accent. However, articulation should be rated with respect to your ability to understand the message, not the social acceptability of the accent. One particular accent is not considered better than another. REMEMBER, in this component you are rating how the student speaks, not what the student says.

ORGANIZATION

The organization rating focuses on how the content of the message is structured. It is concerned with sequence and the relationships among the ideas in the message.

1 = **The organization is inadequate in meeting the requirements of tho task.**

e.g., The message is so disorganized that you cannot understand most of the message.

2 = **The organization is minimal in meeting the requirements of tha task.**

e.g., The organization of the message is mixed up; it jumps back and forth.
The organization of the message appears random or rambling.
You have difficulty understanding the sequence and relationships among the ideas in the message. You have to make some assumptions about the sequence and relationships of ideas.
You cannot put the ideas in the message into an outline.

3 = **The organization is adequate in meeting the requirements of the task.**

e.g., The message is organized.
You do not have difficulty understanding the sequence and relationships among the ideas in the message. You do not have to make assumptions about the sequence and relationships of ideas.
You can put the ideas in the message into an outline.

4 = **The organization is superior in meeting the requirements of the task.**

e.g., The message is overtly organized.
The speaker helps you understand the sequence and relationships of ideas by using organizational aids such as announcing the topic, previewing the organization, using transitions and summarizing.

NOTE: Make sure you are not unconsciously "filling in" organization for a speaker, because you happen to know something about the speaker's topic. If you have to make assumptions about the organization, this fact should be reflected in your rating. REMEMBER, in this component you are rating how the student organizes the message, not what the student says.

CONTENT

The content rating focuses on the specific things which are said. It is concerned with the amount of content related to the task, the relevance of the content to the task and the adaptation of the content to the listener and the situation.

1 = The content is inadequate in meeting the requirements of the task.

e.g., The speaker says practically nothing.
The speaker focuses primarily on irrelevant content.
The speaker is highly egocentric. The speaker appears to ignore the listener and the situation.

2 = The content is minimal in meeting the requirements of the task.

e.g., The speaker does not provide enough content to meet the requirements of the task.
The speaker includes some irrelevant content. The speaker wanders off the topic.
The speaker adapts poorly to the listener and the situation. The speaker uses words and concepts which are inappropriate for the knowledge and experiences of the listener (e.g., slang, jargon, technical language). The speaker uses arguments which are self-centered rather than other-centered.

3 = The content is adequate in meeting the requirements of the task.

e.g., The speaker provides enough content to meet the requirements of the task.
The speaker focuses primarily on relevant content. The speaker sticks to the topic.
The speaker adapts the content in a general way to the listener and the situation. The speaker uses words and concepts which are appropriate for the knowledge and experience of a general audience. The speaker uses arguments which are adapted to a general audience.

4 = The content is superior in meeting the requirements of the task.

e.g., The speaker provides a variety of types of content appropriate for the task, such as generalizations, details, examples and various forms of evidence.
The speaker adapts the content in a specific way to the listener and situation. The speaker takes into account the specific knowledge and experience of the listener, adds explanations as necessary and refers to the listener's experience. The speaker uses arguments which are adapted to the values and motivations of the specific listener.

NOTE: This rating is concerned with content in terms of quantity, relevance and adaptation. It is not concerned with content in terms of accuracy. Concerns with accuracy of content fall outside a speaking skills assessment. Also, make sure you are not unconsciously "filling in" content for a speaker because you happen to know something about the speaker's topic. If you add information, this fact should be reflected in your rating. REMEMBER, in this component you are rating the quantity, relevance and adaptation of what the student says, not the accuracy of what the student says.

LANGUAGE

The language rating deals with the language which is used to convey the message. It is concerned with grammar and choice of words.

1 = **The language is inadequate in meeting the requirements of the task.**

e.g., The grammar and vocabulary are so poor that you cannot understand most of the message.

2 = **The language is minimal in meeting the requirements of the task.**

e.g., The speaker makes many grammatical mistakes.
The speaker uses very simplistic, bland language. The speaker uses a "restricted code," a style of communication characterized by simple grammatical structure and concrete vocabulary.

3 = **The language is adequate in meeting the requirements of the task.**

e.g., The speaker makes few grammatical mistakes.
The speaker uses language which is appropriate for the task, e.g., descriptive language when describing, clear and concise language when giving information and explaining, persuasive language when persuading. The speaker uses an "elaborated code," a style of communication characterized by complex grammatical structure and abstract vocabulary.

4 = **The language is superior in meeting the requirements of the task.**

e.g., The speaker makes very few grammatical mistakes.
The speaker uses language in highly effective ways to emphasize or enhance the meaning of the message. As appropriate to the task, the speaker uses a variety of language techniques such as vivid language, emotional language, humor, imagery, metaphor, simile.

NOTE: In language you may be concerned with students who come from backgrounds where a foreign language or a non-standard form of English is spoken. However, language should be rated with respect to your ability to understand the message, not the social acceptability of the communication style. If a speaker's use of incorrect or non-standard English grammar interferes with your ability to understand the message, this fact should be reflected in your rating. REMEMBER, in this component you are rating how the student conveys the message through language, not what the student says.

APPENDIX B

APPENDIX B

MASSACHUSETTS DEPARTMENT OF EDUCATION ASSESSMENT OF BASIC SKILLS

ONE-ON-ONE SPEAKING TASKS

Description Task:

Think about your favorite class or extracurricular activity in your school. Describe to me everything you can about it so that I will know a lot about it. (How about something like a school subject, a club or a sports program.)

Emergency Task:

Imagine that you are home alone and you smell smoke. You call the fire department and I answer your call. Talk to me as if you were talking on the telephone. Tell me everything I would need to know to get help to you. (Talk directly to me; begin by saying hello.)

Sequence Task:

Think about something you know how to cook. Explain to me step by step how to make it. (How about something like popcorn, a sandwich or eggs.)

Persuasion Task:

Think about one change you would like to see made in your school, like a change in rules or procedures. Imagine I am the principal of your school. Try to convince me that the school should make this change. (How about something like a change in the rules about hall passes or the procedures for enrolling in courses.)