DOCUMENT RESUME

ED 135 838 TM 006 063

TITLE The Annual Conference on Large-Scale Assessment: Formal Papers and Selected Bibliography (Sixth, Boulder, Colorado, June 14-17, 1976).
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
SPONS AGENCY National Center for Education Statistics (DHEW), Washington, D.C.
PUB DATE [Jun 76]
NOTE 139p.

EDRS PRICE MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS Academic Achievement; Agency Role; College Entrance Examinations; *Conference Reports; *Educational Assessment; Elementary Secondary Education; Followup Studies; Hypothesis Testing; Information Utilization; Item Sampling; Kindergarten; Mathematics; Measurement Techniques; Needs Assessment; Performance Tests; Questionnaires; School Districts; Skill Development; Standards; State Agencies; State Departments of Education; *State Programs; Testing Problems; Testing Programs
IDENTIFIERS AAHPER Cooperative Health Education Test; ACT Assessment Program; Delaware Educational Assessment Program; Iowa Assessment Program; Michigan Educational Assessment Program; *National Assessment of Educational Progress; Nebraska Assessment Battery Essential Learn Skills; Pennsylvania Educational Quality Assessment

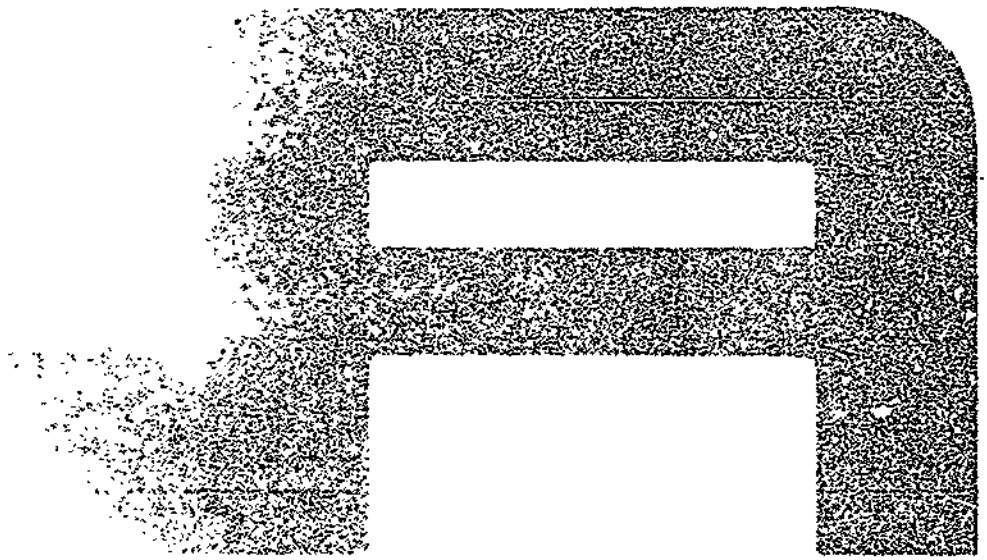
ABSTRACT

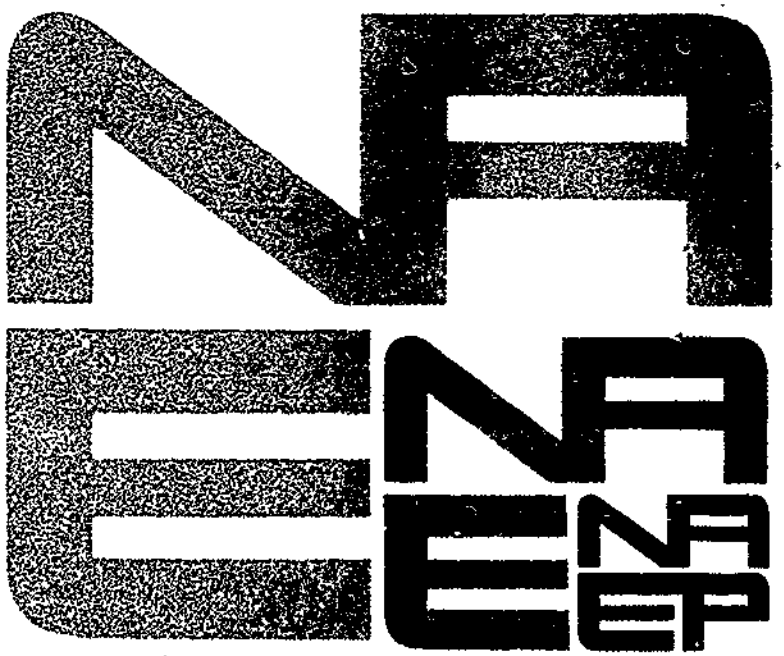
For the past six years the National Assessment of Educational Progress has sponsored a national Conference on Large-Scale Assessment, designed to promote and improve communications among educational assessment personnel in State Departments of Education and other agencies. This volume contains most of the papers that were accepted for presentation at the half-day formal paper session. The 11 papers included here are: (1) "The State Agency as a Resource in Local Needs Assessment" by Paula T. Brictson; (2) "Establishing Criterion Levels for Judging the Acceptability of Assessment Results" by Iris Weiss and Larry Conaway; (3) "N-Abels--A Manageable Technique for Monitoring the Acquisition of Essential Learning Skills" by Harriet A. Egertson and Hugh A. Harlan; (4) "A Process for Developing, Implementing and Following Through on an Assessment Program in Fifth- and Eighth-Grade Mathematics" by Max Morrison; (5) "Educational Quality Assessment Follow-Up Survey of the 1974 Assessment" by Joyce S. Kim; (6) "Hypothesis-Testing in Large-Scale Assessment" by Frank W. Olivas; (7) "A Plan for Utilization of Assessment Data by Local Education Agencies" by John A. Jones and Charles D. Oviatt; (8) "ACT Test Data and Program Assessment for Large School Districts" by Robert Cramer; (9) An Example of the Use of Multiple Matrix Sampling Procedures in a Local District Assessment Program" by Carl D. Novak; (10) "Measurement Problems and Issues Related to Applied Performance Testing" by James R. Sanders; and (11) "Symposium on: Large-Scale Assessment Reporting and Usage: Delaware and Georgia as Exemplars" by Robert Bigelow and Hervey Scudder. (Author)

The Sixth Annual Conference on LARGE-SCALE ASSESSMENT

TM

Formal Papers and Selected Bibliography





NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

The Sixth Annual Conference
on

LARGE-SCALE ASSESSMENT

FORMAL PAPERS AND SELECTED BIBLIOGRAPHY

June 14-17, 1976

Harvest House Hotel

Boulder, Colorado

Sponsored by

Department of Field Services

National Assessment of Educational Progress

A Project of the Education Commission of the States

Funded by

The National Center for Education Statistics
Department of Health, Education and Welfare

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
Suite 700, 1860 Lincoln Street
Denver, Colorado 80203

Roy H. Forbes, *Director*

Conference Co-Directors

Frank B. Womer, University of Michigan

Irvin J. Lehmann, Michigan State University

Jack G. Schmidt, National Assessment

This report has been produced using quality recycled 20 lb. bond paper. For additional copies, write to the address above.

PREFACE

For the past six years the National Assessment of Educational Progress has sponsored a national Conference on Large-Scale Assessment. Designed to promote and improve communications among educational assessment personnel in State Departments of Education and other agencies, the conference has experienced a steady growth in interest, as evidenced by the number of attendees which rose from 17 in 1971 to nearly 200 in 1976.

With this growth have come other changes, mainly in the format and in the substance of the programs. Earlier programs were quite informal, permitting attendees to become familiar with basic features of other state programs and to share with colleagues their experiences and frustrations. More recently the program has become more structured, with additional topics being covered each year and with more attendees taking active roles in formal presentations or leading discussion groups.

The 1976 program, held in Boulder, Colorado, June 14-17, added yet another feature to this evolving conference. For the first time, a half-day of the three-day program was devoted exclusively to presentation of formal papers on a variety of assessment topics. Invitations to submit papers for this portion of the program were extended to previous attendees, various university personnel and to others who had expressed interest in problems of educational assessment. It was anticipated that this volunteer paper presentation program would stimulate assessment personnel to begin documenting some of the innovative and useful procedures that have emerged from current assessment activities.

This volume, Formal Papers and Selected Bibliography, contains most of the papers that were accepted for presentation at the half-day formal paper session. Except for standardizing type size for titles, and making minor changes in formatting, the papers appear as they were submitted by the authors.

Papers were reviewed by a panel of readers, chaired by Jim Impara, formerly of the Oregon assessment program and now a faculty member at Virginia Polytechnic Institute. Readers included Dave Bayless of the Research Triangle Institute, Don Searls of National Assessment, Lorrie Shepard of the University of Colorado, Gordon Ascher, formerly of the New Jersey assessment program and now with the Oregon Department of Education, Bill Burson of the California assessment program and John Adams, formerly of the Minnesota assessment program and now with the Council of Chief State School Officers. To each of them goes a sincere expression of gratitude for their contributions in helping make the paper sessions a very successful first effort.

Authors of submitted papers also deserve our thanks for their work, for without papers and without their permission to publish them, this document would not be possible.

Finally, I wish to use this opportunity to thank two persons who, with me, served as co-directors for the overall conference program -- Frank Womer of the University of Michigan and Irv Lehmann of Michigan State University. Their leadership in guiding this and previous conferences has been of immeasurable value.

JACK G. SCHMIDT, *Director*
Department of Field Services
National Assessment of Educational Progress

TABLE OF CONTENTS

THE STATE AGENCY AS A RESOURCE IN LOCAL NEEDS ASSESSMENT

PAULA T. BRICTSON
MICHIGAN DEPARTMENT OF EDUCATION

Mrs. Paula T. Brictson
Research Consultant
State Department of Education
Box 420
Lansing, Michigan 48902

THE STATE AGENCY AS A RESOURCE IN LOCAL NEEDS ASSESSMENT

PAULA TISS BRICTSON

MICHIGAN DEPARTMENT OF EDUCATION

Abstract

A study in 75 volunteer kindergarten classrooms, designed by a state department of education, allows participating teachers the option of using as many of the state assessment instruments as desired during the entire kindergarten year as one of four assessment modes. Teachers also can use commercial or teacher-made tests, observation, and other adults as information sources about student skill attainment. Each teacher is given a complete list of preprimary objectives, record-keeping forms and copies of the state assessment instruments. Through the state agency as a resource center, a teacher can tailor an assessment program to address the needs and abilities of students within the framework of her/his instructional program.

INTRODUCTION

Several years ago, Michigan's State Board of Education adopted a six step educational management system. The six steps of the system are: the identification of common goals, the development of performance objectives, the assessment of educational needs, the analysis of delivery systems, the evaluation and testing of improvements in these systems, and the development of recommendations for educational improvement. Common goals for the state were approved in 1971, and a series of develop-

mental and review procedures were begun in order to develop statewide minimal performance objectives in the basic skills areas. By 1973, the State Board of Education had given final approval to several sets of performance objectives, one of which was a set of performance objectives for preprimary education. These objectives had been drafted by expert referent groups throughout the state; and reviewed, modified and approved by commissions composed of teachers, curriculum specialists, administrators, and lay citizens.

The Michigan Educational Assessment Program (MEAP) since January, 1970, through the testing of all 4th and 7th graders in basic reading and mathematics skills, has provided information which can contribute to the assessment of needs. The program has collected, analyzed, and disseminated information on district and school resources, student background, school and student academic performance in the basic skills, and school and district size. These data are useful in describing certain aspects of Michigan education and can be used by decision-makers at the school, district, and state levels.

In the school year 1973-74, the educational assessment program made a revolutionary change from norm-referenced to objective-referenced tests based on the several sets of performance objectives. A major change was also made in 1974-75 -- namely, the introduction of a statewide first grade pilot, which was continued with different objectives in 1975-76.

FIRST GRADE EDUCATIONAL ASSESSMENT PILOT

The performance measures administered in the program are designed to test some of the skills of first graders in the affective, cognitive

any psychomotor domains. These skills are considered important for a child to attain before entering first grade. The complete set of The Tentative Objectives for Preprimary Education in Michigan on which the MEAP tests are based is shown in APPENDIX A.

The procedures for the development, validation and editing of the objective-referenced tests used in the 1974 and 1975 First Grade Assessment programs were described in detail in two reports: "Development and Validation of Objective-Referenced Test Instruments for Entry-Level First Grade Children."[1] Briefly, educators from four Michigan school districts (Detroit, Gwinn, Pontiac and Waterford) wrote test items. These items were edited by American Institutes for Research and tried out in the four school districts in two sets. Following each tryout, the items were thoroughly reviewed and revised.

The first grade component is different from the fourth and seventh grade components becuase of the unique requirements of administering objective-referenced tests to students of this age. The tests must be administered either individually or in small groups. This could require enormous amounts of teacher time unless restrictions are placed on the number of items used per objective, the number of objectives tested, and the type of data output expected. This problem was solved by gathering only enough data to yield reliable statewide results and by limiting the number of objectives to be assessed in the program.

In the first year of the first grade educational assessment program (1974-75), 44 individually- and group-administered tests were constructed to measure 48 of the 134 preprimary objectives. Each test was taken by a

---

[1] American Institutes for Research, Palo Alto, California, August, 1974, and June, 1975.

statewide sample of first graders. Approximately 2,500 teachers and 77,000 students were involved in the program. In 1975-76, an additional 32 small group- and individually-administered objective-referenced tests were administered to a statewide sample of 65,000 students. No student selected for the sample was tested with more than one test form.

Teacher feedback elicited through a questionnaire enclosed in the 1974-75 test package indicated that the majority of the behaviors described in the preprimary objectives had already been acquired by entering first graders. The statewide results for the first grade educational assessment program confirmed this teacher observation in that 75% or more of the students correctly answered every test item for 29 of the 48 objectives. In addition, teachers' comments about the usefulness of the information and needed improvements were requested. Some comments were: 1) the information was useful and assessment of preprimary objectives should continue; 2) some teachers suggested that since the majority of entering first grade students had acquired the described behaviors, an educational assessment of preprimary skills would be appropriate at the kindergarten level; 3) a further suggestion was to allow testing over a longer period of time than the three weeks of the regular program.

1975-76 MEAP KINDERGARTEN STUDY

Objectives of the Study

The MEAP kindergarten study has provided the Michigan Department of Education the opportunity to assist 75 volunteer teachers at the classroom level in implementing an educational needs assessment to aid in instructional planning. Teachers were allowed during the period of September,

1975, through April, 1976, to: 1) select from the set of 132 state approved preprimary objectives those important to her/his educational program, 2) assess student attainment of the objectives at an appropriate time in the teaching sequence, 3) choose among four assessment modes a preferred way to test student attainment, and 4) maintain a record of individual and group skill attainment. Thus a teacher can individually design an assessment program to address the needs and abilities of students within the framework of the instructional program.

The desired outcomes for the state agency were to ascertain 1) those preprimary educational objectives important to teachers of kindergarten children, 2) the preferred assessment modes for the numerous preprimary educational objectives, 3) the number of educational objectives which can be assessed during the school year, 4) teacher reaction to the provided test instruments, and 5) teacher reaction to the assessment model prescribed by the study.

Methods

The 1975-76 MEAP kindergarten study evolved from the suggestions of teachers involved in the first year of the first grade pilot program. It seemed appropriate to conduct this study in kindergarten classrooms since many kindergarten teachers have built curricula based upon objectives similar to the set of preprimary objectives and are already assessing student attainment of many of these skills. Rather than limit the assessment of preprimary objectives to only the objectives tested in the test form, the Department asked that teachers focus on the entire set of preprimary objectives.

The Department suggested guidelines for this process and provided, as one type of materials, the 75 MEAP assessment instruments for teachers to use as they felt were appropriate. Teachers could also use three other assessment modes: 1) other assessment instruments (commercial or teacher made tests), 2) teacher observation, 3) and other sources of information about a child's skill attainment, such as another teacher, or parent. The study extended until the end of April, 1976, allowing teachers sufficient time to assess students on skill attainment at a rate which is compatible with each student's development.

During the 1975 spring assessment briefings, the Department staff requested volunteers for the kindergarten special study. Those schools indicating an interest were invited to a briefing which was held in Lansing, Michigan. Following this meeting, district superintendents representing 125 schools sent a letter of commitment to the Department. Schools were stratified according to size and geographical location. A total of 75 classrooms representing 37 districts and 70 schools were selected for this study.

Each teacher received the 75 state test forms, an assessment administration manual for each test, student booklets, and any additional required test materials (beads, bean bags, cassettes, picture books, and so forth). In addition, each teachers was provided an explanatory manual describing the study, a class roster (record-keeping form), directions for recording results and teacher comment sheets. During on-site visits in September, all participating teachers were instructed by a Department staff member on the use of the materials and the parameters of the study.

A class roster (APPENDIX B) for recording attainment of objectives was designed for this study. A column for each of the preprimary ob-

jectives appeared on the roster. When a child attained an objective, the teacher was instructed to indicate under the column, and opposite the student's name, the month the objective was attained and the assessment mode.

Teachers were encouraged to assess as many of the preprimary objectives as possible, using a variety of assessment modes; they were not expected to use only the provided tests for a given objective. One of the desired outcomes of the study is to learn the variety of ways a teacher appraises skill attainment. For example, a teacher could test student attainment of an objective using the provided test with some of the class and test another portion of the class by teacher observation (a different assessment mode). In some cases it might be suitable to use only one assessment mode to measure attainment of an objective.

When the student attains an objective, the teacher indicates on the provided class roster the date (only the month) and the assessment mode. The possible assessment modes are coded as A = MEAP test; B = Other tests; C = Teacher Observation; and D = Other, as explained below.

Assessment Mode A. If a MEAP test is used and the objective measured by that form is attained (using a designated criterion level for each test form) the teacher records a letter A and number indicating the month.

Assessment Mode B. Some teachers have utilized other tests to assess student progress in specific skills. Examples of such tests are those used in local or state evaluation activities, commercial tests, district tests, or their own paper-pencil tests. This study gives the teacher the option of utilizing these tests at her/his discretion. The criterion level for attainment of each objective is determined by the test used. If this assessment mode is used and it indicates the student has attained the ob-

jective, the teacher is instructed to record the letter B and the month the teacher determines the skill is attained.

Assessment Mode C. Teachers may, in the course of teaching, observe students demonstrating attainment of some of the performance objectives. The purpose of providing this assessment mode is to allow teachers who observe attainment to note this and therefore to omit formal testing of these skills for the students observed. In some cases, teachers may design a structured situation, perhaps similar to that used in a more formal test, in order to quickly assess students.

Assessment Mode D. If the teacher judges that a student has developed a specific skill through another assessment mode, such as a student interview or by talking with the student's parents or some technique other than MEAP tests, other tests, or teacher observation, the teacher records the letter D and the month this determination is made.

A copy of the class roster (the record of student attainment) was returned to the Department in May, 1976. Teachers can use their copy for possible use as a diagnostic report to first grade teachers. The rosters will be analyzed by Measurement Research Center/Westinghouse Learning Corporation, Iowa City, Iowa, to determine the number of objectives attained per month and the specific assessment mode used for each objective.

Comment sheets were provided and teachers were invited to comment in five specific areas: 1) comments about the study as a helpful curriculum tool, 2) comments about the study as a facilitative means for assessing student progress in skill attainment, 3) comments about the entire set of tests, 4) comments about specific test items, and 5) comments about the recommended criterion levels. If a teacher chooses other tests as an administration mode, they were asked to describe the tests used for each objective.

In addition to the Class Roster and teacher comment sheets, participating teachers and principals received two mailed surveys which probe their reactions to the study as a viable assessment program for the kindergarten level. Specifically teachers are asked if, having participated in the study and as a result of the methodology of the study, they are in a better position to plan instructional programs based on the needs and strengths of children, and do they have more information about each child's progress. Principals are asked if, as a result of the study, are they in a position to make better judgements about student placement and program planning.

Conclusions

While the final assessment results and surveys were only collected in late May and will not be analyzed before July, interchange at follow-up meetings with teachers in November and February has indicated that the study is viewed as a positive and helpful service by the Michigan Department of Education to assist teachers in tailoring a local needs assessment.

In response to the enthusiasm of teachers and principals, the Department will offer the kindergarten program on a volunteer basis in 1976-77. Originally this coming year's project allowed for the inclusion of 100 elementary shcools. Due to the extensive requests to participate, the study will be conducted in 200 schools. The Department will provide each participating school a manual containing 1) test administration and scoring directions for each MEAP test, 2) individual and group record keeping forms, 3) suggested observation techniques, and 4) example classroom activities to assess student skill attainment. In addition, a kit containing all handouts (cassettes, picture books, beads, geometric shapes, alphabet cards, and so forth) for administering MEAP tests will be included. A set of

ditto masters of the student booklets will be provided. This should alleviate storage problems and permit the teacher the option of duplicating only those specific tests which he or she decides to use. Regional meetings will be help in August and September to brief participants on the details of the program.

Implications of the Study

It is critical for state agencies, in addition to providing curriculum and program guidance, to instruct and assist local education agencies in the design and use of assessment techniques for local assessment programs. Further, the state agency can serve a resource service by providing tests to measure a variety of curriculum objectives important to local educational programs. The study is proving to be a viable model whereby a state agency can provide an assessment program design, tests, and record-keeping forms, and yet permit autonomy at the local level.

APPENDIX A

AFFECTIVE OBJECTIVES FOR PREPRIMARY STUDENTS

A. EMOTIONAL BEHAVIOR

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Recognize at least three of five basic emotions (fear, anger, sadness, joy, love) in self and others;
2. Recognize some basic causes of familiar emotional responses (e.g., sad, happy, angry, etc.);
3. Begin to show empathy for and awareness of the feelings, needs, and desires of others;
4. Actively express feelings nonverbally;
5. A greater ability to verbalize affective experiences (e.g., positive and negative feelings, wants, values, conflicts, etc.);
6. Display an increased repertoire of behavioral responses by which to solve affective problems (e.g., create their own solutions; seek help from parents, teachers, and others; give help to other children; etc.);
7. Given situations in which gratification must be delayed, will demonstrate increased ability to accept imposed delay and to regulate behavior appropriately.

B. SELF CONCEPT

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. An increase in positive self-image;
2. Given role-playing and real-life situations, will demonstrate an increased awareness of their relationship to their family and to the wider community and environment;
3. Given role-playing and real-life situations, will demonstrate an increased awareness of racial and cultural similarities and differences;
4. An increased understanding of the concept of sexuality (i.e., recognize their sexual identification; are comfortable with own sexuality and the sexuality of others);
5. Given role-playing and real-life situations, will demonstrate a healthy, self-respecting attitude towards their bodies and its simple physiological functions;
6. Given various roles to play (such as occupational, parental, emotional, cultural, or situational) will demonstrate awareness and sensitivity for these roles.

C. SOCIAL RELATIONSHIPS

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Widen peer and adult relationships by demonstrating increased ability to play with one or more children and to relate to a larger group;
   - 1.1 An increased capacity to cope with strange and/or new surroundings and with familiar and unfamiliar people;
   - 1.2 An increased ability to seek help from others when needed and when appropriate;
2. Begin developing social interdependence by exhibiting an increased awareness of the importance of give-and-take in social and work relationships;
   - 2.1 Exhibit evidence that they are accepting of differences in others;
   - 2.2 Demonstrate their ability to listen to others;
   - 2.3 Exhibit the quality of sharing with others;
   - 2.4 Demonstrate that they have learned to ask permission to use objects belonging to another person;
   - 2.5 Demonstrate that they can recognize cause and effect in the behavior of others, and the effects of their behavior on others;

2.6 Exhibit greater participation in activities and in communication with others;
3. Identify several workers from different occupational areas in the community and tell something about their work;
4. Name some of the people children learn from and what they learn from them;
5. Participate in decision-making situations (e.g., make personal or group rules for classroom behavior, etc.).

D. BEHAVIORAL RESPONSE TO CLASSROOM ENVIRONMENT

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Willingness to accept reasonable limits set upon behavior, play space, use of materials, or the type of activities in which engaged;
2. Acceptance of routines (e.g., daily schedules, room arrangements, adults, etc.) and changes in routines;
3. Cooperation and independence (without help or demonstration) in following verbal directions for three or more sequential instructions;
4. Increased independence in the areas of personal hygiene, eating, and dressing;
5. Increased ability to independently begin, work through, and continue an activity;
6. Increased ability to accept responsibility for the use and care of their portion of the classroom environment.

PSYCHOMOTOR OBJECTIVES FOR PREPRIMARY STUDENTS

A. GROSS MOTOR BEHAVIOR

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Balance while walking (e.g., will be able to walk at least ten feet on a straight three-inch taped line without stepping completely off the line with either foot);
2. Balance while running (e.g., will be able to run to a target placed no more than twenty feet away without stopping or veering off a path approximately five feet wide);
3. Muscle coordination (e.g., will be able to jump with both feet rising together over a three-inch taped line);
4. Muscle coordination and balance (e.g., will be able to hop three consecutive times using one foot);
5. Eye-foot muscle coordination and balance (e.g., will be able to kick a ten-inch ball without losing his balance or falling);
6. Eye-hand coordination (e.g., given a bushel basket tilted toward him at a 45-degree angle and placed four feet in front of him, the child will throw a bean bag into the basket);
7. Touch or move parts of the body (e.g., head, arms, elbows, hands, legs, knees, feet) called for by the teacher;
8. Free body movement by physically responding to music, song, rhythm, and/or rhymes;
9. Leg coordination (e.g., will be able to skip or gallop, leading with the preferred foot).

B. FINE MOTOR BEHAVIOR

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Digital coordination (e.g., place a three quarter inch button through a one-inch button hole);
2. Digital coordination (e.g., by being able to place ten small one-half inch beads on a lacing string);

3. Eye-hand coordination (e.g., given a ten-minute time limit, will be able to put together a simple puzzle of five to eight pieces);
4. Thumb-finger coordination (e.g., given a pair of child's scissors and a strip of one-inch by six-inch construction paper, can make clean cuts three times in five attempts without folding or tearing the paper);
5. Eye-hand coordination (e.g., given a large crayon and at least a two-inch model of a circle, will be able to copy the model in such a manner that the curved line closes);
6. Eye-hand coordination and lateral movement (e.g., given a large crayon and at least a two-inch model of two intersecting lines, will be able to copy the lines so that they intersect in some manner);
7. Improved eye-hand coordination (e.g., given materials such as interlocking blocks or other available small blocks, will be able to build a stable eight-piece vertical structure or design).

COGNITIVE OBJECTIVES FOR PREPRIMARY STUDENTS

A. LANGUAGE DEVELOPMENT

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Enjoyment in looking at books and listening to stories;
2. Produce pictures and/or scribbles of own creation which are used as a basis for communication;
3. Listen and react to another's oral language;
4. Given an oral story which expresses a mood (e.g., happy, sad, angry, afraid), will identify the characteristic mood of the story;
5. Given an oral stimulus requiring a specific bodily response (e.g., the game "Simon Says"), will provide the appropriate response;
6. Talk about a picture or a group of two or three related pictures;
7. Tell about personal experiences;
8. Distinguish environmental sounds they hear (e.g., traffic sounds, dog barking, baby crying, etc.);
9. Given three single syllable sounds, two of which rhyme, will select the two which rhyme;
10. Express an idea or ask a question orally of another person (e.g., explaining how a toy works/asking how a toy works);
11. Given a small group situation, will share own ideas and listen to the ideas of others;
12. Talk about the feelings associated with events;
13. Non-verbally imitate or role-play the simple action of people or animals;
14. Name likenesses and differences in pictures, objects, and shapes;
15. Recognize some letters of the alphabet;
16. Given a sequence of pictures portraying a story, will tell about the story by responding appropriately to each picture;
17. Print first name correctly;
18. Recognize first name.

B. CLASSIFICATION AND ORDERING

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Given two kinds of objects in a large set (e.g., elbow and shell macaroni or bottle caps and checkers), will sort the objects into two sets according to their separate characteristics;
2. Given an object of a specific color, will pick an object which is of the same color;

3. Group items on the basis of common function (e.g., things to eat with, things to wear, things to play with, etc.);
4. Group items on the basis of association (e.g., hammer and nail, shoe and foot, etc.);
5. Identify and group items on the basis of general classes or categories (such as furniture, animals, plants, etc.);
6. Given items of common qualities (e.g., texture, weight, loudness, speed, temperature, color), will group and match items on the basis of these qualities and be expected to know and use at least two of the comparative terms (e.g., soft-hard, loud-quiet, fast-slow, smooth-rough, hot-cold, dark-light, heavy-light) to identify the groupings;
7. Given a pattern using objects of two or more colors, will duplicate the pattern selecting from a set of similar objects;
8. Given a set of ten objects of assorted color and shape, will pick out objects having specific combinations of the two attributes;
9. Given one series of three objects arranged in a pattern by color or shape and the first object of the second series, will complete the second pattern series;
10. Given a variety of objects, will group some of the objects into a classification system according to their own perceptions.

C. NUMBER — NUMERATION

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Given a set of coins of a penny, nickel, dime, will pick and name each one;
2. Given a collection of five objects of varying lengths, will pick up the *longest* or the *shortest* as requested.
3. Given a set of five pictures of objects of various heights, will arrange the pictures so that the objects are ordered from shortest to tallest;
4. Given two objects of decidedly different weights, will hand to the teacher the one that is heavy or the one that is light as requested;
5. Given the direction "count to ten", will recite the number names from one through ten in the usual order;
6. Given an oral description of a set and a collection of objects, some of which belong to the set and some of which do not, will pick up the objects that are members of the given set;
7. Given cutout pictures of any two sets (from one to five members), will place the pictures of the sets in order, from that set with less members to that set with more members; then, will order the set pictures from more to less;
8. Given numeral cards 1 through 5 and five sets of objects consisting of one, two, three, four and five members, will place the sets in sequential order from the set with fewest to the set with the most and then will place the numeral cards in front of the set having the number of members named by the numeral;
9. Given a set of objects with 1-5 members, will count the members of the set and state the cardinal number of that set;
10. Given pictures of sets with 0-9 objects and number cards from 0-9 (using felt numerals, sandpaper numerals), will match the right numeral with the picture of the set having the same number of members;
11. Given dot pattern cards showing sets of 0-10 dots, will count while pointing to the appropriate dot card;
12. Given a set of 2 to 8 objects, the students, from his own group of more than 8 objects will construct a set having more members than the original set;
13. Given a set of 2 to 8 objects, the students, from his own group of objects will construct a set having fewer members than the original set;

14. Given an assortment of cutout shapes including squares, triangles, rectangles and circles of various sizes randomly arranged, will select a given shape as requested.

D. SPATIAL RELATIONS

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Identify and name the following parts of his body: head, arms, hands, torso, legs and feet;
2. Knowledge of concepts of position (such as on-off, over-under, on top of, in-out, into-out of, top-bottom, above-below, in front of-in back of, behind, beside-next to, by, between);
3. Knowledge of concepts of direction (such as up-down, around-through, forward-backward, to-from, sideways, across);
4. Knowledge of concepts of distance (such as near-far, close to-far from).

E. TEMPORAL RELATIONS

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective-referenced instruments:

1. Ability to follow temporal commands (such as go, stop, at the same time, now, start, finish);
2. Understanding of time intervals (such as beginning-end, fast-slow).

F. NATURAL SCIENCES

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Given objects of various primary colors (red, blue and yellow), will be able to correctly identify the colors;
2. Given an object to examine using their senses of sight, sound, smell, taste, and touch, will exhibit the ability to describe certain characteristics (such as size, color, weight, texture, temperature, odor, etc.);
3. Given an object (or picture of an object), will describe verbally by naming at least two characteristics of the object (e.g., given a rubber ball, the student will give two of the properties, such as color, shape (round), density (light), elasticity (bouncy), size (smaller than my hand), temperature (cool), texture (smooth);
4. Given a set of objects or events, will arrange them in sequence in accordance with prescribed criteria (e.g., given separate pictures of a dog and a puppy or a flower and some seeds, the student will arrange them in proper order);
5. Given an object or picture which changes with successive observations, will state at least one of the properties which is changing (e.g., the student tastes a sample of unbaked cookie dough and a sample of a cookie made from the same dough and describes what changed in the baking (hardness, texture, color, taste, smell);
6. Given a magnifying glass and an object or organism with some characteristic not visible without a lens, can observe the object or specimen with the lens and identify at least one of the characteristics;
7. Given a picture or group of pictures showing items which comprise both live and non-live things, can point to examples of living and non-living things.

G. SAFETY

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Awareness of common hazards encountered in daily living (e.g., toxic household chemicals or substances, electricity, toxic plants, explosive and combustible substances, etc.);

2. Adhere to safety rules in the home, to and from school, and in the school;
3. Perform safely as pedestrians, as passengers in motor vehicles, and as tricycle operators.

H. FINE ARTS

Art:

The joy in creativity should be emphasized throughout all fine arts instruction. The process is more important than the product. By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Pleasure and enjoyment in a variety of art experiences;
2. Use a variety of media (such as paint, crayons, finger paint, felt markers, etc.);
3. Create two- and three-dimensional forms using a variety of manipulative materials (such as clay, paper-mache, blocks, etc.);
4. Recognize color in the natural environment and in the man-made environment;
5. Use a variety of color in the production of art;
6. Recognize that lines define space (e.g., uses line in a variety of ways to express length, size, or shape);
7. Recognize the direction of line (e.g., down, slanted, over, across, etc.);
8. Identify the characteristics of line (e.g., fat, thin, winding, climbing, etc.);
9. Use a variety of lines in his art activities;
10. Distinguish between two- and three-dimensional forms;
11. Develop compositions using size, shape, direction, overlapping shapes and/or repetition;
12. Use a combination of various textures in art forms;
13. Recognize differences in his art work (e.g., size, surface, parts of objects, shape, texture, etc.);
14. Use flat, curved and irregular surfaces in producing three-dimensional forms.

Music:

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Create music on a variety of classroom instruments;
2. Freely express the mood of music through body movement;
3. Through physical movements (e.g., clap, march, walk, run, play rhythm instrument) demonstrate his ability to respond rhythmically to pulse or beat in music;
4. Repeat a very simple rhythm, individually or in a group (e.g., singing, chanting, speaking, clapping, using rhythm instruments);
5. Participate with a group in singing simple, familiar melodies;
6. Upon hearing music, will recognize whether a melody moves up or down;
7. Upon hearing music, will recognize fast and slow tempos;
8. Distinguish between long and short tones.

I. AESTHETIC APPRECIATION

By the end of the preprimary experience, students should be able to demonstrate the following behaviors as measured by teacher observation and/or objective referenced instruments:

1. Begin to develop aesthetic appreciation by responding emotionally, through non-directed, spontaneous self-expression (drawing, painting, movement, self-report), to moods and feelings in art, music, movement, drama, poetry, prose and nature;
2. Begin to recognize the beauty or aesthetic qualities of his own work as well as the work of others;

3. Value his art experience (e.g., feels comfortable with art activities, willingly participates in art activities, expresses personal satisfaction with art activities, voluntarily elects to repeat the art experiences, demonstrates pride in art work, expresses himself through color, etc.);
4. During an art activity, will voluntarily use a variety of patterns and both two- and three-dimensional forms;
5. Indicate a preference for certain textures in the daily art experience;
6. React to musical experience by voluntarily responding in out-of-school situations (e.g., discusses music class happenings, sings songs learned at school, chooses to listen to music programs on radio, television, etc.);
7. React to musical experience by voluntarily responding during school (e.g., expresses a reaction when it is time for music, joins in quickly, freely, or slowly when musical activities begin, expresses reactions to the music class during classtime or when it has ended, brings a favorite record to school, seeks opportunities to play classroom instruments, etc.).

APPENDIX B

CLASS ROSTER

ASSESSMENT MODE
A = MEAP Tests
B = Other Tests
C = Teacher Observation
D = Other

Teacher _____

School _____

District _____

AFFECTIVE DOMAIN

		Emotional Behavior							Self Concept						Social Relationships			
STUDENT NAME	MEAP Test No.	1101	1102	X	X	X	X	X	X	X	X	X	X	X	1301	X	X	X
	Prepimary Obj.	1	2	3	4	5	6	7	1	2	3	4	5	6	1	1.1	1.2	2.1
1																		
2																		
3																		
4																		
5																		
6																		
7																		
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		
25																		
26																		
27																		
28																		
29																		
30																		

ESTABLISHING CRITERION LEVELS FOR JUDGING THE ACCEPTABILITY OF ASSESSMENT RESULTS

IRIS WEISS AND LARRY CONAWAY

LARRY CONAWAY, PRESENTER

RESEARCH TRIANGLE INSTITUTE

Mrs. Iris R. Weiss
Education Research Scientist
Research Triangle Institute
PO Box 12194
Research Triangle Park, N.C. 27709

Mr. Larry Conaway
Senior Educational Research Scientist
Research Triangle Institute
PO Box 12194
Research Triangle Park, N.C. 27709

ESTABLISHING CRITERION LEVELS FOR JUDGING THE ACCEPTABILITY OF ASSESSMENT RESULTS

IRIS WEISS and LARRY CONAWAY

RESEARCH TRAINGLE INSTITUTE

I. INTRODUCTION

A. The Problem

One approach for judging the acceptability of student performance has been to compare state and local assessment results to national and regional results reported by the National Assessment of Educational Progress. These types of comparisons have provided useful information for identifying possible strengths and weaknesses in curriculum and instruction at the state and local levels. However, educators who have been involved in interpreting these results have indicated a need for additional types of criteria. The fact that the state performance levels on certain items were significantly below the national performance levels does not necessarily mean that these are areas of weakness in the state's program. It may be that these items are poorly constructed items or that they are measuring skills which are considered to be of low priority in that state. Similarly, if the Nation's students performed poorly in a particular area, surpassing the Nation is not necessarily an indication of strength. A state can do better than the Nation but still do a poor job. It was obvious that these types of statistical comparisons of performance would not suffice; professional judgment would have to be brought to bear in identifying strengths and weaknesses.

Several states and local districts have attempted to obtain this professional judgment by involving educators in interpreting assessment results. They have brought together groups of educators to study the assessment results, to determine areas where performance was particularly weak, and to make recommendations for change. Unfortunately, the criteria which they used to identify strengths and weaknesses were often quite vague. With the actual results in front of them, educators found it difficult to think about what they really would have liked performance to be. There was a tendency to consider areas with low p-values as weak and those with high p-values as strong without regard to the inherent difficulty of the items used to assess these areas. There was also a tendency to "make excuses" for poor performances, to

judge a performance level acceptable because it was "about the best you could expect considering the level of present instruction." The idea that instruction might therefore need changing sometimes did not emerge.

There was a need for procedures to establish á priori standards with which educators could compare student performance to assist them in judging the acceptability of assessment results. This paper describes procedures which have been developed to establish á priori standards and discusses some of the uses and limitations of these procedures.

B. The Literature

The review of the related literature provides only a little guidance in establishing practical procedures for setting á priori performance standards. Airasian and Madaus (1972) stated that the area of setting standards was the area of criterion referenced measurement in most need of research, and Quirk (1974) discussed a similar lack of research in the context of performance based teacher education.

Two very recent articles contended that the lack of research in this area still exists today. Meskauskas (1976) reviewed procedures suggested for setting the pass-fail point and concluded that only one method has received extensive practical use. This method was developed by Nedelsky in 1954 and involves teacher judgment of the distractors which minimally passing students will be able to identify as incorrect. This model has been used extensively in medical education, and literature is available on several applications including a recent one by Smilansky and Guerin (1976).

Jaeger (1976, p. 13) also reviewed a number of standard setting procedures, and he summarized the present situation as follows:

> The research that exists on standard-setting procedures in competency-based education appears to be largely theoretical .... Throughout this paper, I have identified questions for which research-based answers are apparently unavailable. Application of statistical models and theoretical formulation are unlikely to provide answers to these questions. There is need instead for empirical investigation involving human standard-setters in real or simulated judgmental situations, using real performance data and real descriptions of task domains.

After a decision was made to develop a new set of procedures for setting á priori standards, the literature did provide some guidance in deciding to have educators set standards for individual items rather than objectives. Millman (1973) stated that it is difficult to defend the frequent practice of employing a particular passing score only on the grounds of tradition. Quirk (1974, p. 317) was very specific in his discussion of fixed cutoff scores in performance based teacher education:

> While [fixed cutoff scores] sound semiscientific, they do not possess much substantive value. The percentage of items related to an objective which a candidate answers correctly is a function not only of the content of the items, but also of the difficulty of the items. An estimate of the difficulty of the items can be obtained either from a logical judgment based on a study of the specific items or from empirical item-analysis data.

After discussing the state of research in standard-setting, the state-of-the-art in domain referenced testing, and the problems associated with learning hierarchies, Airasian and Madaus (1972) concluded that teachers will have to establish their own standards using expert opinion, experience, face validity of items, and group consensus.

C. Development and Use of the Procedures

The procedures were first developed in early 1974 by staff members of the Research Triangle Institute, the Minnesota Department of Education, and the University of Minnesota. They were first used in conjunction with the Minnesota statewide assessment of reading and the Maine statewide assessment of reading by surveying a sample of teachers with mailout questionnaires. Consensus procedures were then developed, and these have been used by Research Triangle Institute staff in the Richfield (Minnesota) local school district, the Guilford (Connecticut) local school district, and the Maine and Washington statewide assessments. Similar procedures have also been used by others in conjunction with the Oregon statewide assessment, and the Minnesota Educational Assessment Program has continued to use these procedures.

Those who have used these procedures have not attempted to present or use the á priori criterion standards as absolutes. Rather, the standards have been presented as carefully considered professional judgments. They have been used along with other information--normative data and individual item results--to assist those who interpret assessment results in judging the acceptability of these results.

II. OBTAINING A PRIORI ESTIMATES OF MINIMAL ACCEPTABLE, DESIRED, AND PREDICTED PERFORMANCE LEVELS

A. Definitions

The procedures developed for establishing criterion levels involved having educators examine individual assessment items and estimate minimal, desired, and predicted performance levels. The definitions were specifically adapted for each assessment, but they have been generally defined as follows:

1. Minimal Acceptable Outcome - The percent of students you believe must be able to respond correctly to a particular item in order for you to consider instruction to be providing essential skills to these students.
2. Desired Outcome - The percent of students you believe should be able to respond correctly to a particular item.
3. Predicted Outcome - The percent of students you believe will respond correctly to a particular item.

B. Statewide Samples

In the two earliest studies statewide samples of third and fourth grade teachers in Minnesota and Maine were asked to make these estimates for reading items administered to 9-year-olds using mailout questionnaires. The teachers' responses to these mailout questionnaires were averaged to provide statewide minimal acceptable, desired and predicted performance levels for each item. Actual student performance levels were then compared to each of these levels. In addition, average student performance levels on groups of items in an objective were compared to teacher expectations averaged over these same items. Items or objectives which had actual performance above the desired level could be considered strengths, and those which were below the minimal acceptable levels could be considered weaknesses.

Three independent samples of teachers were involved in each study; the teachers' estimates for each item were relatively stable across the three samples, indicating basic reliability of the estimates obtained using the instrument. The results also showed that the teachers' estimates of predicted performance were generally quite close to actual student

performance; they were within 15 percentage points on 77% of the items in one study and on 73% of the items in the other study. When averaged across items in an objective, the teachers' predictions were extremely close to actual student performance. The procedures and results of these studies are presented in detail in a paper by Elliott (1974).

At first glance these results seem to indicate that teachers are remarkably good predictors of student performance. However, using averages of teacher responses conceals the fact that for some items there was marked variability in teacher estimates. For example, the teachers' predictions for one item averaged to 62.4%, which was extremely close to the actual student performance of 59.9%. However, nearly one-fourth of the teachers predicted that at least 80% of the students would answer this item correctly, while more than 10% of the teachers predicted p-values of 40% or less. Similarly, when obtaining the average prediction for an objective, the fact that the teachers overpredicted on some items and underpredicted on others resulted in predicted average values which were very close to actual average performance values. Clearly, these averages do not indicate whether the teachers are good diagnosticians.

Using averages obtained by mailout survey techniques to determine minimal acceptable and desired criterion levels causes similar problems. For example, in an extreme case, half of the teachers may think an item is inappropriate for all but the brightest 9-year-olds and may therefore set the minimal acceptable level at 10%; the other half of the teachers may think the item is vitally important for 9-year-olds and may set the minimal acceptable level at 90%. Using the average value (50%) as the minimal acceptable criterion level does not really reflect the teachers' judgments.

The analysis problems involved in using averages raised the question of whether some other indication of central tendency such as median or modal response should be used in determining minimal acceptable and desired performance levels. More importantly, the various types of important considerations causing great diversity of opinion on some items pointed out the need for groups of educators to communicate about

the objectives of instruction in a given subject area. For this reason, and because using large samples of teachers was quite expensive, consensus procedures were adopted so that relatively small groups of educators would be able to meet together to establish criterion levels.

C. Consensus Procedures

The first study utilizing consensus procedures with a group of educators was conducted in Richfield, Minnesota, a school district which was conducting a local assessment in conjunction with the Minnesota statewide reading assessment. A committee of educators representing a wide range of classroom situations from remedial to advanced ability groups met to discuss the assessment items and to reach consensus on minimal acceptable, desired, and predicted performance levels.

The consensus procedures yielded more extreme estimates than did averaging responses of a statewide sample of teachers. For example, for identical items administered to 9-year-olds the predicted levels established by Richfield teachers using consensus procedures ranged from 20% to 98% while the statewide estimates ranged only from 47% to 75%. The differences in ranges were similar for minimal acceptable and desired performance levels.

Results are available for comparing predicted levels established by consensus procedures with student performance from three studies. The Richfield, Minnesota assessment of reading was conducted at ages 9 and 13; the Maine assessment of mathematics was conducted at ages 13 and 17; and the Guilford, Connecticut assessment of science was conducted at age 17. These results are shown in Table 1.

Table 1

PERCENT OF ITEMS FOR WHICH THE CONSENSUS PREDICTED PERFORMANCE LEVEL WAS WITHIN 15 PERCENT OF THE ACTUAL STUDENT PERFORMANCE LEVEL

	Richfield		Maine		Guilford
	Age 9	Age 13	Age 13	Age 17	Age 17
0% - 5%	31%	36%	27%	23%	23%
6% - 10%	20%	22%	20%	20%	16%
11% - 15%	19%	13%	11%	21%	16%
TOTAL	70%	71%	58%	64%	55%

In general, the consensus groups have been reasonably accurate in their predictions of student performance. As can be seen in Table 1, across the five consensus groups the percent of items for which predicted performance levels were within 15 percent of actual student performance ranged from 55% to 71%.

The teachers' estimates of minimal acceptable and desired performance levels have been used to determine a priori classifications of strength or weakness in objectives within the subject area. The average student performance level across the items in an objective is compared to the average minimal acceptable performance level and the average desired level across the items. While the cutoff points have varied somewhat for different studies, student performance above the desired level has generally been defined as a strength for an objective, and student performance below the minimal acceptable level has generally been defined as a weakness.

The locally developed criterion levels have been used to assist in judging the acceptability of student performance. The procedures used in obtaining these estimates have yielded beneficial side effects as well; they have provided a vehicle for establishing communication among teachers within a school and across schools within a school district. The discussions have focused attention on educational objectives and student capabilities, and they often result in open debate about what should be taught compared

to what actually is taught. Criterion levels developed through the use of consensus procedures for statewide assessments have proved similarly useful in establishing communication among teachers, administrators, and university and state department of education personnel.

III. INTERPRETING ASSESSMENT RESULTS

A. Some Problems in Interpreting A Priori Criterion Levels

The discussion thus far has focused on procedures for establishing criterion levels prior to examining assessment results. Some of the advantages of these a priori criteria have been pointed out. However these procedures also have some serious limitations. Many teachers who have been involved in establishing statewide criterion levels have indicated that they felt uneasy making statewide estimates; they would have felt much more comfortable setting minimal acceptable, desired, and predicted performance levels for their own classes. The same problem exists in local districts, where some teachers have very slow classes and others teach only the most gifted children, but the extent of the problem is considerably smaller at the district level.

An additional limitation of these procedures is due to the fact that estimates are based on the percent of students who answer each item correctly. The relative frequency of certain types of errors do not enter into the establishment of these a priori criterion levels. This is a serious weakness since it is likely that error patterns would affect judgments about the acceptability of student performance. Consider the following hypothetical situation: The minimal acceptable level for an open-ended mathematics word problem item was set at 60%; only 40% of the students were correct, which may indicate that performance was weak. However, another 35% of the students set up the problem correctly but made minor computational errors. Thus a total of 75% of the students demonstrated that they knew how to go about finding the correct answer, and a group of educators might well judge this performance to be satisfactory.

As another example of error patterns affecting judgments of acceptability consider the following hypothetical example: The desired performance level for an item involving metric measurement was set at 70%, and 69% of the students were correct, which would indicate that performance was satisfactory. However, many students chose a response alternative which indicated that meters are used to measure volume. Many mathematics educators might consider the frequency of this error to show an area of weakness even though the percent correct was at about the desired level.

A related problem is the fact that some prior expectations are based on rather inaccurate judgments about the difficulty of certain items. For example, at first inspection the following item seemed to be quite straightforward, and a committee of mathematics educators predicted that 50% of Maine 17-year-olds would choose the correct answer. The actual performance (32% correct) was well below both the minimal acceptable level (60% correct) and the desired level (75% correct) which would seem to indicate a weakness.

A housewife will pay the lowest price per ounce for rice if she buys it at the store which offers

%	
7.8	12 ounces for 40 cents.
6.6	14 ounces for 45 cents.
32.1	1 pound, 12 ounces for 85 cents.
47.8	2 pounds for 99 cents.
5.9	I don't know.
0.0	Missing

The assessment results showed that a very large number of students chose foil 4 rather than the correct response. The committee examined the item closely and speculated that there were two major reasons for the performance on this item. First, choice 4 was the largest package size, and larger sizes typically have smaller costs per ounce. Second, the correct answer had a price of 3.04 cents per ounce while choice 4 had a price of 3.09 cents per ounce. To answer the item correctly a

student would have to be able to set up the problem correctly and carry out each division to two decimal places; the problem was not one of simple estimation as it had first appeared. The committee concluded that, considering the complexity of the item, student performance was satisfactory.

The procedures described in this paper are based on the belief that the actual assessment items as well as the objectives have to be considered when setting criterion levels since the difficulty of specific items must be considered. However, conclusions of strength or weakness have been based on groups of items in objectives or skill areas rather than on individual items. One of the potential drawbacks of these procedures is that they result in conclusions of strength, satisfactory, or weakness for each objective regardless of the number of items in the objective. These results should be interpreted with great caution when there are only 2 or 3 items in an objective since educators may feel that the objective is inadequately measured and that no conclusion is justified.

B. Combining A Priori Criterion Levels with Professional Judgment of Assessment Results

To avoid some of the problems associated with establishing á priori criterion levels without sacrificing the advantages, attempts were made in Maine and Guilford, Connecticut to formally combine professional judgments of assessment results with conclusions based on the á priori criterion levels. In each case the committee had used consensus procedures to set minimal acceptable and desired criterion levels. The next step was carried out several months later, and the committee members did not have access to these á priori criterion levels. Each committee member was given a student assessment booklet which included indications of the percent of students who chose each response alternative, including "I don't know." Each committee member was asked to rate each item from +2 (highly satisfactory) to -2 (highly unsatisfactory), and then the entire group was brought together to discuss the results and reach consensus on the item ratings. During this meeting there was much more

discussion of error patterns because actual student results were available for each foil or response category. These item ratings were then used to determine whether actual student performance on each objective was strong, satisfactory or weak.

The consensus committees were then able to look at both their á priori classifications of strong, satisfactory, and weak objectives and their ex post facto classifications when interpreting results. It is interesting to compare the results of the á priori and ex post facto procedures. In the Maine statewide mathematics assessment there were several areas identified as weaknesses by both procedures, e.g., percents, fractions and word problems at age 13. Other areas, such as geometry at age 17, were considered to be weak when the educators looked at the actual results although the á priori criteria had resulted in conclusions of satisfactory performance. Conversely, 13-year-old performance in the area of probability was judged satisfactory after the fact even though average performance on the items was well below the minimal acceptable level. In several other cases, the committee decided that there were too few items to draw firm conclusions about performance.

In the Guilford science assessment there were two areas identified as strengths by both procedures--evolution-related items and health-related items. Some areas, such as light and electricity, were classified as strengths by á priori criterion standards but as satisfactory by ex post facto judgments. Generally, however, there was a tendency for ratings to be higher after the fact. For example, some areas, such as biological science, were classified as satisfactory by á priori criterion standards but as strengths by ex post facto judgments. Similarly, there were three areas of weakness according to á priori criterion standards, but no areas of weakness according to ex post facto judgments.

The committee members considered the á priori expectations to be an important stage in the process of determining strengths and weaknesses and interpreting assessment results. Working closely with the items at an early stage in setting á priori criterion levels for the assessment provided for great familiarity with the instruments. Also, comparison of actual performance with desired and minimal acceptable levels of

performance, which were established without actual performance results, provided valuable reference points in judging the acceptability of performance within objectives. Finally, committee members identified individual items for which performance was well below the desired level; they then looked at these items and the types of errors students frequently made and came up with ideas for improving the curriculum.

IV. DIRECTIONS FOR FURTHER DEVELOPMENT

In addition to developing techniques for judging the acceptability of student performance, the use of teacher predictions of student performance for determining needs for in-service staff development and pre-service teacher education are being investigated. For example, predictions can be used to determine if some teachers are poor predictors of student performance while others are very accurate predictors, or if most teachers are quite far off on quite a few of the items. Predictions can also be used to determine if teachers recognize which skills their students have and have not mastered and what types of errors will be most frequent.

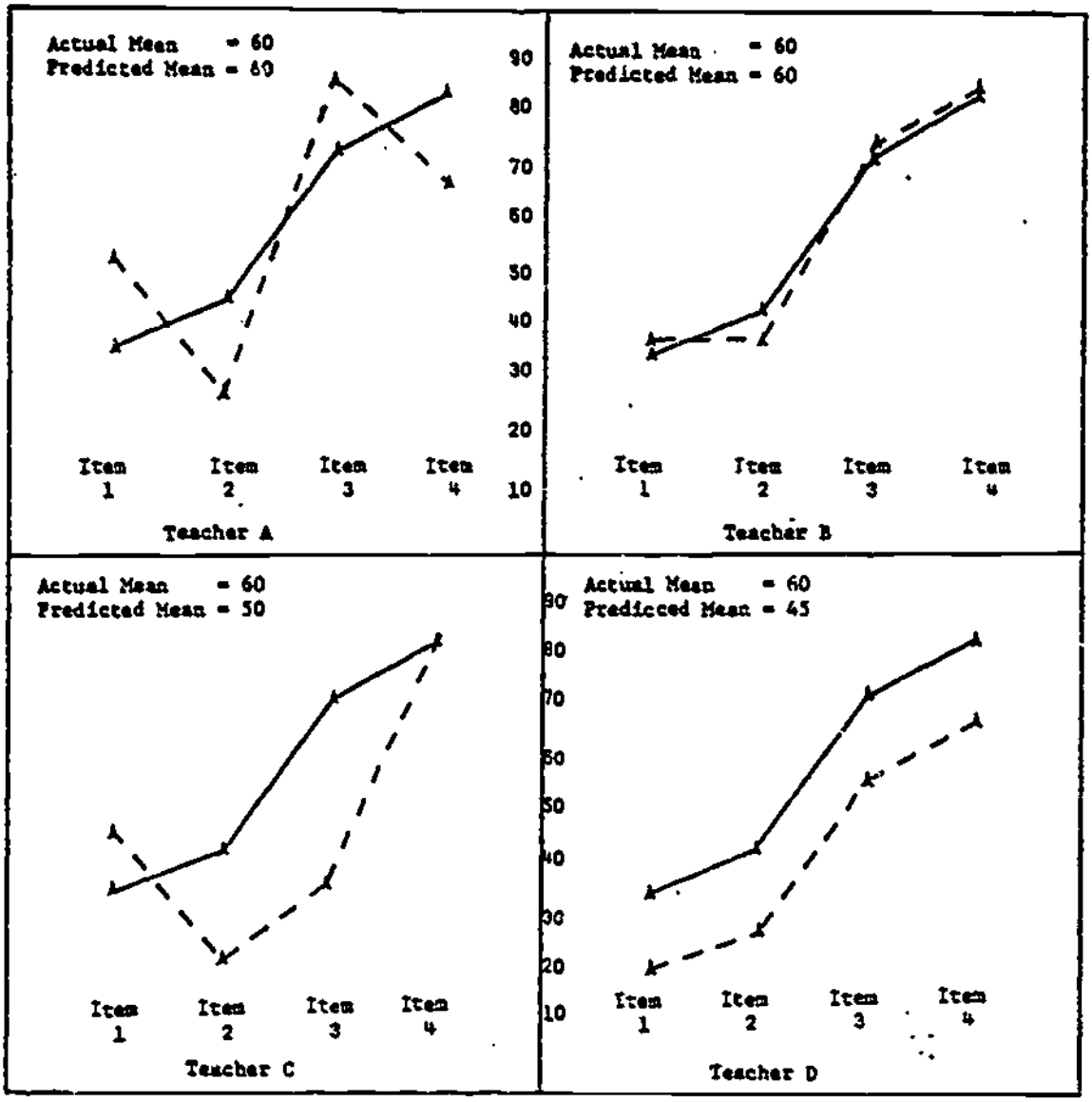
A procedure has been developed which compares teacher predictions and actual performance to determine the extent to which teachers recognize the relative difficulty of items within each objective.[1/] In this procedure, a profile of teacher predictions is compared to a profile of student performance. The overall distance between the two profiles consists of two components: the distance between the means of the two profiles and the residual difference between the two profiles when adjusted for mean differences.

Figure 1 shows hypothetical examples of actual and predicted profiles for a mathematics objective involving fractions. There were four items in this objective; their p-values ranged from 38% to 84% with a mean of 60%. The profile analysis procedures determine the overall distance between the predicted and actual performance profiles. They also determine the amount of this difference which is due to (1) a tendency to either underpredict or overpredict and (2) a failure to recognize the relative difficulties of the four items.

1/ Dr. Anthony J. Conger of RTI developed this procedure by adapting one described in Wiggins, Jerry S. Personality and Prediction: Principles of Personality Assessment, Reading, MA: Addison Welsey, 1973.

Figure 1: COMPARISON OF ACTUAL AND PREDICTED PERFORMANCE PROFILES

Item	Actual	Teacher A	Teacher B	Teacher C	Teacher D
1	38	53	40	50	23
2	45	30	40	25	30
3	73	88	75	40	58
4	84	69	85	85	69



These examples show that teachers A and B had predicted means which were identical to the actual mean, while teachers C and D had predicted means which were 10 and 15 percentage points below the actual mean, respectively. These results do not necessarily mean that teachers A and B are good diagnosticians and teachers C and D are poor diagnosticians. Table 2 shows that, when you adjust for the difference between means, predicted profiles D and B are closest to the actual profile (adjusted distances of 0 and 8.5, respectively). These results indicate that teachers D and B were quite accurate in predicting the relative difficulties of the four items. Teacher C showed the least ability to recognize relative difficulties. For example, he predicted that item 1 would be easier than item 3 when in fact many more students were correct on item 3 than on item 1 (73% versus 38% correct). Teacher ability to recognize relative item difficulties is shown graphically in Figure 1 by the fact that the profiles in example D are parallel, while the profiles in examples A and D are clearly not parallel.

Table 2

ANALYSIS OF THE DIFFERENCES BETWEEN
ACTUAL AND PREDICTED PERFORMANCE PROFILES

Teacher	Predicted Mean	Actual Mean	Actual Mean Minus Predicted Mean	Overall Distance	Mean Distance	Adjusted Distance
A	60	60	0	225.0	0	225.0
B	60	60	0	8.5	0	8.5
C	50	60	10	408.5	100	308.5
D	45	60	15	225.0	225	0

Another area currently being investigated is having teachers predict which errors will be most common so these predictions can be compared to the actual results. These types of results, as well as the results of profile analyses, can be useful for planning teacher education programs. Teachers who are unable to predict performance patterns may also need help in diagnosing student learning needs. If certain objectives seem to pose problems for many teachers, in-service and pre-service programs can focus on techniques for diagnosing student needs and instructional materials for meeting these needs.

Time and budget constraints have almost completely prevented research into the varying performance standards which would be established by different consensus groups working with identical assessment items and populations of students. However, one small study of this problem was done in conjunction with the 1976 Washington statewide assessment of fourth grade mathematics.

The Washington Mathematics Committee was divided into two separate groups of seven members each to establish desired and predicted performance levels for mathematics items. The groups were about equally representative in terms of university mathematics educators, teachers, and administrators. Each group was to establish desired and predicted levels for approximately half of the items, but eight items were assigned to both groups. The groups worked independently after being given the same instructions. For the eight items both sets of criterion standards were announced, and the full committee established final consensus standards.

Table 3 shows the results of the study. While this study was far too small and informal to be conclusive, the results are encouraging. In establishing the desired level for the eight items the two groups had exactly the same levels for two items, and they were as far as 15 percent apart only once. In establishing the predicted level the two groups were never in exact agreement, but they were within five percent twice, and they were as far as 15 percent apart on only three items.

Table 3

DESIRED AND PREDICTED CONSENSUS LEVELS ESTABLISHED BY GROUP A, GROUP B AND THE COMBINED COMMITTEE

Item	Desired Level			Predicted Level		
	Group A	Group B	Combined	Group A	Group B	Combined
1	50	65	60	40	55	50
2	75	75	75	65	70	67
3	75	85	80	60	75	65
4	75	80	80	65	70	70
5	80	90	85	75	85	80
6	80	90	85	70	80	75
7	80	90	85	75	85	80
8	70	70	70	50	65	55

Student performance results are not yet available from the Washington assessment, but it will be interesting to see if one group was consistently better than the other in predicting item performance. It will also be interesting to see if student performance is usually closer to the consensus of the combined group than it is to that of either Group A or Group B.

REFERENCES

Airasian, Peter W. and Madaus, George F. Criterion-Referenced Testing in the Classroom. NCME Measurement in Education, May 1972, Vol. 3, No. 4, 1-8.

Elliott, Muriel C. Teacher Outcomes Studies: The Development of Methods for Obtaining Teacher Estimates of Minimal and Desired Student Performance. Paper Presented at the Southeastern Invitational Conference on Measurement in Education, Knoxville, December, 1974.

Jaeger, Richard M. Measurement Consequences of Selected Standard-Setting Models. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Meskauskas, John A. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard-Setting. Review of Educational Research, 1976, 46, 133-158.

Millman, Jason, Domain-Referenced Measures. Review of Educational Research, 1973, 43, 205-216.

Quirk, Thomas J. Some Measurement Issues in Competency-Based Teacher Education. Phi Delta Kappan, 1974, LV, 316-319.

Smilansky, Jonathon and Guerin, Robert O. Minimal Acceptable Performance Levels for Criterion-Referenced Multiple Choice Examinations and Their Validation. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

N-ABELS -- A MANAGEABLE TECHNIQUE FOR MONITORING THE ACQUISITION OF ESSENTIAL LEARNING SKILLS

HARRIET A. EGERTSON AND HUGH A. HARLAN
HARRIET A. EGERTSON, PRESENTER
NEBRASKA DEPARTMENT OF EDUCATION

Ms. Harriet A. Egertson
Consultant, School Management Svs.
State Department of Education
233 South 10th Street
Lincoln, Nebraska 68508

Mr. Hugh Harlan
Administrator, School Management Svs.
State Department of Education
233 South 10th Street
Lincoln, Nebraska 68508

N-ABELS -- A MANAGEABLE TECHNIQUE FOR MONITORING THE ACQUISITION OF ESSENTIAL LEARNING SKILLS

HARRIET A. EGERTSON and HUGH A. HARLAN

NEBRASKA DEPARTMENT OF EDUCATION

Introduction

In the spring of 1974, the Nebraska Department of Education began to develop an assessment instrument which focused on skills that could be considered essential for continuing success in school and for learning independence. The product of this work was the publication in the summer of 1975 of the Nebraska Assessment Battery of Essential Learning Skills. N-ABELS can be described as a goal-oriented teaching instrument with an evaluation component providing a basis for determining acquisition of twelve defined skills. Each goal is stated in terms of an acceptable performance standard which clearly describes what the student must do.

Over half the schools in Nebraska are using the battery on a voluntary basis this year and early information received is most promising. It is hoped that the use of this battery will have the following effects: 1) to assure the public that their stated priorities are taken seriously by the schools; 2) to help the public accept new programs by assuring mastery of essential skills; 3) to answer requests for accountability without imposing legal prescriptions and restraints; and 4) to clarify the continuing responsibility of each teacher to work toward competency in essential skills for each student.

The Development of N-ABELS

Skill Selection

The twelve skills included in N-ABELS were chosen by committees of professionals in three general areas: communication skills, mathematics

skills, and inquiry skills. A set of predetermined criteria for choosing skills aided the selection process. The skills in N-ABELS are those:

1) For which the school assumes the primary instructional responsibility.

2) Which are necessary for independence in learning.

3) Which engender wide public agreement concerning their importance.

4) Which are commonly introduced in the elementary school.

5) Which can be assessed by readily demonstrable student performance.

6) Which can be assessed without prescribing teaching methodology.

Each skill is defined in terms of what the student must do to demonstrate mastery and the actual tests were constructed to conform to these definitions. The test is at the same difficulty level on all forms (Sample and Forms A-D are available). Because these skills are considered important for a student to be successful in school, they should be acquired by students as soon as possible. Of course, most students will achieve far beyond this level. The primary purpose of the test is to pinpoint individual students who still need help in learning essential skills. Assessment begins in the upper elementary grades providing ample time for aiding students with problems in a particular skill. Once assessment is initiated, a student continues working on the skills until all of them have been mastered. This may take several years. As soon as students demonstrate mastery in any one of the twelve skill tests, they are finished with that portion and are not retested on that skill.

Communication Skills

All of the tests in N-ABELS which assess communication skills are constructed from a vocabulary list of 2000 words that was compiled during the course of the development of N-ABELS. The decision to use a

standard vocabulary list of most frequently used words as a way of defining communication skills made it possible to develop tests which assess the actual spelling and reading knowledge of a finite set of words.

Two recent computer studies showed some promise of utility for this project. Both concentrate on materials written for children. The American Heritage Word Frequency list completed in 1971 was prepared for use as the basis for the new American Heritage School Dictionary (Carroll, 1971). It is based on 1,045 five hundred word samples (a total of more than 5 million running words) from a variety of text and trade materials for grades three through nine. Every unique symbol which appeared in the samples was included so that the list includes single letters, numerals, proper nouns, inflected forms, and formulae. The words are listed in rank order according to frequency of occurrence.

Basic Elementary Reading Vocabularies (Harris, 1972) is also a computerized list of selected materials for children; however, it differs from the American Heritage list in several ways. The 4,500,000 running words of this list are the entire vocabularies of six basal reader series, and two series each in English, mathematics, science, and social studies. Three types of lists are provided: 1) core lists contain words which occur in a majority of the textual material at each grade level; 2) additional lists contain words which occur less frequently at each level; and 3) technical lists of more difficult words from the subject areas of English, social science, science, and math. Inflected forms were not included except for first and second grade. The 7,613 words are arranged alphabetically by grade level so that frequency is impossible to determine.

Since neither of these lists precisely fit the needs of N-ABELS, it was decided to do a computerized comparison based on the American Heritage list. The two lists were consolidated and the N-ABELS list was determined using a set of predetermined criteria.

Three skill tests are provided in the communication area. The reading skill test requires that the student demonstrate the ability to translate printed symbols into speech by reading aloud a narrative selection of approximately 100 words constructed from the Nebraska-Assessment Battery of Essential Learning Skills 2000 Word Reading Vocabulary List.

In the area of writing two tests are provided. The writing skill test is based on the first 1000 words of the N-ABELS Vocabulary and requires the student to write legibly, spell correctly, and punctuate appropriately from dictation a 100 word selection. Criteria for determining legibility and correct punctuation are provided. In addition, a supplementary spelling skills test requires the student to spell correctly 20 of the more difficult words randomly selected from the N-ABELS 1000 Word Writing Vocabulary.

Mathematics Skills

The forty-eight Mathematical Competencies and Skills Essential for Enlightened Citizens developed by the National Council of Teachers of Mathematics was used as a source for the math skills listed in N-ABELS. (Edwards, Nichols, and Sharpe, 1972) The NCTM competencies and skills, however, include the full range of mathematics knowledge deemed necessary for enlightened cititzenship in a person's life beyond the school. The skills included in this composite are those which fit the general purposes of N-ABELS, that is, the skills which students need to know to be able to progress independently in school.

The math skills tests require the student to demonstrate the ability to read and write positive rational numbers correctly; to associate positive rational numbers using decimal, percent, and the fractional notation for halves, thirds, fourths, fifths, eighths, and tenths with concrete or pictorial representations of objects; to write the basic facts sums and products from a dictated tape; and to use the four standard operations of arithmetic for whole numbers and decimal fractions.

Inquiry Skills

The purpose of assessing the inquiry skills listed in N-ABELS is to demonstrate the student's ability to operationally use the inquiry tools cited. There is no effort to assess the ability to understand or analyze any material discovered in reference sources. The assessment of such higher level cognitive skills is beyond the scope intended in this instrument.

The inquiry skills tests require the student to demonstrate the ability to locate words in a dictionary, to locate topics in an encyclopedia and extend that search to cross-reference items, to use the card catalog to find materials on a given topic, and to use the current Official Highway Map of Nebraska to locate places using cardinal directions, identify physical and political features on a map, and to estimate the distance between locations using the map scale. As much as possible, the tests for these skills involve demonstration of the ability to efficiently locate information in the sources themselves.

Administering N-ABELS

Although N-ABELS is intended to be used primarily on an individual basis, it is possible to administer some of the tests in large groups.

This is particularly helpful when the process is first initiated. Retesting is generally done on an individual basis. The format is so flexible that schools have been able to adapt the administration procedures to their particular situations without compromising the intent of the instrument.

Our field testing indicated that most of the skill tests can be administered using materials available in the classroom with little expense to the school district. It was also established that the tests can be administered on a continuing basis with minimal disruption of the regular school program.

Some of the test procedures are unfamiliar; therefore, many students will not demonstrate mastery the first time a skill test is taken. The errors made should be reviewed with the student and practice activities should be suggested which will help the student gain competence in this skill. Each student should be encouraged to keep working toward mastery of a specified skill which the teacher feels is most likely attainable. If the student is near mastery, retesting should be scheduled within one or two weeks. In addition to the school record, each student should have a copy of the Student Progress Report Form so that a record of successes can be shared with parents and friends.

References

Carroll, J. B., Davies, P. & Richman, B., The American Heritage Word Frequency Book. New York: Houghton Mifflin, 1971.

Dale, E. & Reichert, D., Bibliography of Vocabulary Studies. Columbus, Ohio: Bureau of Educational Research, Ohio State University, 1957.

Edwards, E. L. Jr., Nichols, E. D. & Sharpe, G. H., Mathematical Competencies and Skills Essential for Enlightened Citizens, The Arithmetic Teacher, XVIX (November, 1972), 601-607.

Harris, A. J., & Jacobson, M. D., Basic Elementary Reading Vocabularies. London: Macmillan, 1972.

Horn, E. A Basic Writing Vocabulary, University of Iowa Monographs in Education, Series 1, No. 4 (April, 1926), 3-226.

A PROCESS FOR DEVELOPING, IMPLEMENTING AND FOLLOWING THROUGH ON AN ASSESSMENT PROGRAM IN FIFTH- AND EIGHTH-GRADE MATHEMATICS

MAX MORRISON
IOWA DEPARTMENT OF EDUCATION

Dr. Max Morrison
Director:PRE
Department of Public Instruction
Grimes State Office Building
Des Moines, Iowa 50125

A PROCESS FOR DEVELOPING, IMPLEMENTING AND FOLLOWING THROUGH ON AN ASSESSMENT PROGRAM IN FIFTH- AND EIGHTH-GRADE MATHEMATICS

MAX MORRISON

IOWA DEPARTMENT OF EDUCATION

(Summary)

Because of a growing concern over current measuring instruments and in response to criticisms regarding student achievement in mathematics, members of the Iowa Council of Teachers of Mathematics (ICTM) and staff from the Iowa Department of Public Instruction initiated a program for statewide mathematics assessment in June 1974. It was designed to collect pertinent and specific data on student achievement that could be used by teachers in diagnosing and prescribing instruction. To assist the Department staff in developing and implementing the program, a nine member committee consisting of classroom teachers, mathematics supervisors and college mathematics instructors was established. The committee's initial role was to establish criteria for an effective assessment program and to monitor progress.

The goals of statewide mathematics assessment were:

1. To provide specific information on each student which could be used by the teacher to diagnose each student's strengths and weaknesses in mathematics;

2. To provide objective data for each teacher to furnish the basis for planning sequential learning activities for the entire class or for each individual in the class;

3. To provide data for school districts that could be used in revising the curriculum and in planning inservice activities for the staff;

4. To provide benchmark information to the Iowa Department of Public Instruction so that performance trends over time can be studied; and

5. To provide a process which could be replicated by local school districts in determining the effectiveness of each curricular offering.

After the committee reviewed procedures used by National Assessment in identifying objectives and test items and the techniques used to sample students and items, the committee developed the following criteria for the Iowa Assessment program:

1. Participation by local schools in the assessment program would be on a voluntary basis;

2. A list of minimal performance objectives would be identified to insure that important objectives are not overlooked;

3. Student testing would be limited to the cognitive domain with only grades five and eight included.

4. Four items would be developed to measure the attainment of each objective;

5. The test would differ from a norm-referenced test in that it would not be designed for the purpose of comparing one student's performance with that of another;

6. The entire test battery would be administered to each student so that the data could be used for diagnostic purposes;

7. Test items would be developed which would incorporate the use of recall, application and analysis skills;

8. The test would be administered early in the school year to enable the teacher to utilize the data in planning instructional activities throughout the year;

9. Test scoring would be the responsibility of the classroom teacher with the results reported to the student as soon after testing as possible;

10. Each student's performance would be recorded on a single page profile sheet to be developed by the committee;

11. Assessment items would be limited to those which could be measured by paper and pencil;

12. The major focus of the assessment would be to provide the teacher with data for making decisions on individual students; and

13. Collection of the data at the state level would be for the purpose of identifying problems common to a number of schools and to provide baseline data which would be used to study performance trends over time.

The state assessment program did not attempt to measure proficiency on all desirable mathematics skills and concepts. Many worthwhile experiences such as constructing geometric figures using compass and ruler or estimating the length, height, or weight of objects in the classroom were not included.

A set of minimal objectives was identified for beginning fifth and eighth grade students following an extensive survey of current textbook content and after reviewing mathematics objectives identified by other states and those identified in National Assessment. Iowa's objectives were based upon skills and concepts deemed essential for future success in mathematics, or skills required to deal with solving practical problems in everyday life situations. The first list of objectives developed were submitted to 150 mathematics teachers throughout the state for comments regarding their appropriateness. Revisions were made and items were developed to measure student performance on each objective. These items were pilot tested in four school districts and the test was then revised.

Local school districts were invited to participate in mathematics assessment in March 1975. Requests for participation came from 140 school districts throughout the state resulting in the distribution of some 20,000 fifth grade tests and 22,000 eighth grade tests.

Mathematics teachers from the participating schools were asked to review the objectives and test items prior to administering the test. A part of the review included getting expected levels of performance for each class. Each teacher was asked to estimate the per cent of students he/she believed would demonstrate mastery on each objective. This information would then be useful to the teacher in analyzing the results as he/she could compare the actual performance against the expected level of performance.

To assure comparability of data, local teachers were requested to administer the assessment tests between September 15, and October 17, 1975. Teachers were to score the tests and record the results on individual profile sheets. Individual profiles were to be recorded on a class profile or school profile and forwarded to the Iowa Department of Public Instruction where a state profile was to be developed. As participation was voluntary, schools who did not elect to send in class or school profiles were not contacted to submit a report.

A unique feature of the Iowa Assessment Program which distinguishes it from programs in other states is the assistance provided to the teacher prior to and following the assessment. Area education agency consultants and local school math coordinators arranged pre- and post-test assessment activities with local teachers. For example, a consultant would schedule a meeting of the staff to validate the objectives. Teachers would compare local objectives against those included in the assessment. Consultative assistance was also available to assist the teacher in analyzing the results and to determine appropriate instructional activities.

The State Mathematics Committee developed the following aids: an assessment handbook to explain the how and why of assessment; a guide for diagnosing errors on the fifth grade test; cassette tapes of all fifth and eighth grade test items for individual administration; and a list of suggested instructional activities for the measurement strand focusing on objectives included in the assessment. Other guides are to be developed upon the request of the teachers.

Comments

Contrary to many of the recent criticisms that students lack basic math skills, the results of the Iowa Assessment show that a large majority of students have acquired a good foundation in mathematics. Evidence of success can be noted on computation of whole numbers in both fifth and eighth grade. The per cent of success declined somewhat on computation of fractions, decimals and percentages, but students should have an opportunity to further develop these skills during the remainder of the school year.

Performance of eighth grade students on word problems reveal that slightly more than one-half of the students were able to apply previously learned skills when confronted with a problem solving situation. The lower rate of success may be partially attributed to the type of word problems used in the assessment. A number of problems required the student to analyze the information carefully in order to discard the extraneous data prior to solving the problem. The student's previous experience with this type of problem may have been extremely limited.

Results indicate that when students are confronted with a word problem where they are required to estimate or approximate a reasonable answer, more than 50 percent of the students are unable to select the "best estimate". Math programs have not stressed this skill in the past, but with the increased dependence upon pocket calculators and other automatic calculating equipment, it becomes much more crucial to be able to judge the reasonableness of an answer.

One could speculate further on the results of state assessment, but the crucial judgments regarding the use of the data must be made by the local school staff. Some questions that should be raised by the teachers as a result of the assessment include:

1) Was the overall performance of students satisfactory?
2) Which students did not perform satisfactorily?
3) What were their skill deficiencies?
4) How serious are these deficiencies?
5) What are the consequences if nothing is done about correcting the deficiencies?
6) What resources are available to assist with the problem?
7) How long will it take to resolve the problem?
8) What action can be taken to prevent similar problems from occurring?
9) What skill maintenance activities are necessary for all students?
10) What other objectives should be included in the assessment?

Seeking answers to the above questions or a similar set developed by the teachers should enable schools to pinpoint the problem and to allocate resources to resolve the situation.

EDUCATIONAL QUALITY ASSESSMENT
FOLLOW-UP SURVEY OF THE 1974 ASSESSMENT

JOYCE S. KIM
PENNSYLVANIA DEPARTMENT OF EDUCATION

Dr. Joyce S. Kim
Educational Research Associate
State Department of Education
Box 911
Harrisburg, Pennsylvania 17126

EDUCATIONAL QUALITY ASSESSMENT
FOLLOW-UP SURVEY OF THE 1974 ASSESSMENT

JOYCE S. KIM

PENNSYLVANIA DEPARTMENT OF EDUCATION

Introduction

On November 9, 1973, the State Board of Education adopted Section 5.76, Educational Quality Assessment as follows:

> "During the school years 1973-74, 1974-75 and 1975-76, the Department of Education shall use the Educational Quality Assessment procedure to evaluate the effectiveness of the educational program for all school districts in the state based upon the Ten Goals of Quality Education adopted by the State Board of Education. Public schools housing approximately one-third of the students enrolled in each of the three grades 5, 8 and 11 will be included in the assessment each year."

Approximately one-third of the districts (170 districts) participated in the 1973-74 assessment undertaken by the Division of Educational Quality Assessment (EQA). The districts assessed during the 1973-74 school year contained:

Grade	Number of Schools	Number of Students
5	785	51,342
8	240	53,326
11	191	48,276
TOTAL	1,216	152,944

All of the participating districts received their school reports from the Division of Educational Quality Assessment in the fall of 1974.

OBJECTIVES OF THE STUDY

EQA is designed to offer reliable, statistical information on strengths and weaknesses, from which schools can base sound educational planning decisions. Schools are free to use the results as they wish. EQA's function is to provide schools with the starting point for a self-analysis of their programs.

The EQA follow-up survey was carried out in order to ascertain what effect the data and information disseminated by the Division of EQA, have had on the local school programs. Some of the questions answered were: To what extent have the assessment results been disseminated? What value do school districts see in EQA? How relevant are the EQA results to educational and planning decisions?

PROCEDURES -- METHODS AND TECHNIQUES

The follow-up survey was conducted in all districts that had participated in the March 1974 assessment. Careful consideration was given in selecting one-third of the districts (170 districts) for the assessment. The criteria for selection of a representative sample were: size of the district, socioeconomic level determined by the financial aid ratio and geographic balance.

In October 1975, the Division of EQA mailed a 20-item Follow-up Opinionnaire to superintendents of 170 districts. Replies were received from 138 districts by the end of December 1975. As shown in the Appendix, the Opinionnaire included important questions for both schools and EQA to ascertain reactions to the assessment. Questions in the opinionnaire were

focused on 1) extent of dissemination of EQA results, 2) usefulness of EQA data and 3) contribution of the EQA advisory service. Survey results are valuable for EQA because the Division is in the process of revising procedures, measurement instruments and condition variables for post-1976 assessment.

RESULTS

Through the follow-up survey, it was found that a wide dissemination of the results had been made. The assessment results have been disseminated to various categories of publics and organizations. The approximate number of persons involved in each category were:

Persons Involved	Number
School board members	974
Principals	766
Central office staff	511
Most elementary teachers	9,327
Most middle/junior high school teachers	5,375
Most high school teachers	7,074
Local service clubs (Lion, Jaycees, etc.)	596
PTA, PTO, any parent group	7,574
Students	37,891
General public	390,536
Other: Citizen Advisory Committee	3,533
Community Advisory Council	
Counselors	
In-service for staff	
Long Range Advisory Group	
Middle States Association	
Newspaper	
Psychological Interns	
Superintendent Lay Advisory Council	
Taxpayer Association	
University Consultants, etc.	

N = 111 districts

This dissemination included many different approaches -- written reports, newsletters, in-service presentations, school board meetings, etc. In addition, 82 per cent of the districts had prepared press releases for their local newspapers. Therefore, it is evident that EQA information is being shared with public and is not buried in a school administrator's desk drawer. Different methods were used to inform others about the EQA report. They include:

QUESTION: WHAT METHODS HAVE BEEN USED TO INFORM OTHERS ABOUT THE EQA REPORT?

Methods Used	Per Cent
Special written report	57.2
School district newsletter	51.4
Press release	81.9
Faculty Memorandum	34.1
Curriculum Bulletin	13.8
In-service presentations	75.4
School board meeting	90.6
Faculty meeting	89.9
PTA presentation	45.7
Special presentation	34.8
Regular meeting	21.0
Other	7.2
None	.7

N = 138

Over 50 per cent of the districts indicated that new programs are being established as the result of EQA information. About 93 per cent of the districts claimed that EQA information has been reviewed with building administrators in program planning. Only 6.5 per cent of the districts said that they had not, as yet, used the EQA data. Of these districts, most stated that non-use of data was due to lack of time on the part of district personnel. Only one district felt that the information was not sufficiently credible to merit use. Therefore, one can conclude that use

of EQA data is definitely being made even though use of the data is not mandated by the Department of Education.

QUESTION: WHICH OF THE FOLLOWING DESCRIBE THE USE MADE OF THE EQA INFORMATION?

Use Made	Per Cent
The information has not, as yet, been used..	6.5
The information was used to reflect the favorability of our present programs.....	51.4
A new program is being planned for one or more of our schools as a result of the information.....	51.4
Revisions of some existing programs are underway as a result of the information.....	60.1
The information has been reviewed with building administrators for their use in program planning.....	92.8
The information served as a basis for teacher in-service activity.....	58.7
A new program has been "tried out" in one of our schools as a result of the information.....	9.4
A new program has been incorporated into one school's program as a result of the information.....	17.4
A new program has been incorporated into several of our schools as a result of the information.....	18.1

N = 138

Most school districts (93.5%) were satisfied with the EQA interpreting team's report and more than two-third of the districts indicated that they do not need a follow-up interpretation session.

As far as goal priorities are concerned, the particular goals often chosen as the five most useful in planning (out of ten) were: Basic Skills-Verbal, Basic Skills-Math, Self-Esteem, Citizenship and Interest in School and Learning. Each of the basic skills received the highest vote (54% of the districts). This means that about 46 per cent of the districts did not include basic skills in the top five. This is probably due to the fact that they already have information in the basic skills area which they can use in program planning.

QUESTION: MARK THE FIVE GOAL AREAS FOR WHICH THE INFORMATION PROVIDED BY THE ASSESSMENT WAS MOST USEFUL IN PLANNING.

Goal Areas	Per Cent
Self Esteem	51.4
Understanding Others	34.8
Basic Skills-Verbal	54.3
Basic Skills-Math	54.3
Interest in School and Learning	44.2
Citizenship	40.6
Health	28.3
Vocational Attitude	24.6
Vocational Knowledge	29.0
Creative Activities	29.7
Appreciating Human Accomplishments	23.2
Preparing for a Changing World	24.6

N = 138

More than one-half of the districts administered standardized achievement tests (1973-74 school year) at grade levels two through eight. Only two out of 138 districts said they had not given any standardized achievement tests. However, this testing had been done in other school years. Those two districts explained that they declared a moratorium on testing during the 1973-74 school year.

QUESTION: AT WHICH GRADE LEVEL(S) DID YOU GIVE A STANDARDIZED ACHIEVEMENT TEST IN THE 1973-74 SCHOOL YEAR?

Grade Level	Per Cent	Grade Level	Per Cent
1	48.6	7	59.4
2	63.8	8	58.7
3	73.2	9	41.3
4	71.0	10	26.8
5	74.6	11	31.9
6	83.3	12	15.9
		NONE	1.4

N = 138

Most districts felt that the EQA date have the most relevance for change in teaching strategies. Many (more than 80 per cent) also felt it was relevant in changing course offerings and course content.

QUESTION: HOW RELEVANT IS THE INFORMATION PROVIDED IN THE REPORT TO DECISIONS WHICH MUST BE MADE IN THE FOLLOWING AREAS?

Changes Made In:	Very Relevant	Quite Relevant	Relevant	Not Relevant	No Answer
Course offerings	5.8%	26.8%	47.8%	15.2%	4.3%
Course content	8.7	35.5	43.5	9.4	2.9
Teaching strategies	12.3	37.0	36.2	7.2	7.2
Teaching assignments	1.4	3.6	29.0	58.0	8.0
Financial allocations	3.6	10.9	44.2	34.8	6.5
School facilities	1.4	9.4	23.9	59.4	5.8

N = 138

Some districts entertained suspicions of existing problems, but they did not have data to support those suspicions. Over 83 per cent of the districts reported that EQA had either provided them the data to confirm such suspicions, or EQA data called attention to problem areas which were not previously detected by district staff.

THE EQA INFORMATION:

18.1% -- a) called attention to a problem area not previously noted by district staff.
52.2% -- b) confirmed suspicions about district problems.
13.8% -- c) did not identify any serious problems.
2.9% -- No response.
10.9% -- Combination of a and b.
2.2% -- Combination of a, b and c.

N = 138

It is significant to observe that 81 per cent of the districts considered the EQA program as a means of helping them make decisions, and about two-thirds of the districts believed that the EQA information represents a true picture of their district.

QUESTION: HOW DO YOU CONSIDER THE EQA PROGRAM AS A MEANS OF HELPING YOU MAKE DECISIONS?

15.5% -- a) Very useful
65.6% -- b) Useful
15.6% -- c) Not very useful
1.7% -- d) Useless
1.4% -- e) No response

N = 138

QUESTION: DO YOU BELIEVE THE EQA INFORMATION REPRESENTS A TRUE PICTURE OF YOUR DISTRICT IN THE AREAS ASSESSED?

64.5% -- Yes
22.5% -- No
13.0% -- No response

N = 138

About 65 per cent of the districts said they used criterion-reference subscale scores in their explanation of the results, and more than three-fourth felt that item response frequency data would be valuable if given with the initial report.

QUESTION: DID YOU USE CRITERION-REFERENCE SUBSCALE SCORES IN YOUR EXPLANATION OF THE RESULTS?

64.5% -- Yes
31.2% -- No
4.3% -- No response

N = 138

QUESTION: WOULD ITEM RESPONSE FREQUENCY DATA BE VALUABLE IF GIVEN WITH THE INITIAL REPORT?

75.4% -- Yes
17.4% -- No
7.2% -- No response

N = 138

Condition variables which were collected to identify the differences in resources among schools came primarily from students and teachers as part of their assessment questionnaires. Seventy-one per cent of the districts claimed that condition variable data is not harmful. Less than four per cent said that teacher response variables and home, education and occupation variables are harmful. Another 71 per cent expressed that the follow-up booklets on suggested strategies are valuable.

QUESTION: IS CONDITION VARIABLE DATA MORE HARMFUL THAN HELPFUL?

10.9% -- Yes
71.0% -- No
18.1% -- No response

N = 138

QUESTION: WERE THE FOLLOW-UP BOOKLETS ON SUGGESTED STRATEGIES OF ANY VALUE?

71.0% -- Yes
4.3% -- No
13.8% -- Not received
10.9% -- No response

N = 138

Of the 20 items included in the Opinionnaire, 4 questions ask about suggestions, comments and evaluations regarding EQA's program. These questions are:

- What do you feel is the value of EQA?
- Can you suggest any programs, approaches or techniques used in your district that may account for any high scores you have? (Above the 75th percentile or above the prediction band)
- What type of assistance would you like from the EQA as a result of the assessment?
- Please add any comments you wish regarding the assessment programs.

Value of EQA

In general, most districts felt that EQA has provided a valuable tool for accountability and evaluation to measure the Ten Goals of Quality Education. Over two-thirds of the districts said that EQA provides objective measurement of general effectiveness of school operations and allows an overall picture of the extent to which students are achieving the goals. Twenty-two out of 138 districts (15.9%) mentioned EQA's value in terms of measurement for affective domains of education. EQA's ability to measure affective goals which can not be measured easily, especially relative to state norms is remarked by districts. Seven per cent of the districts said that the instruments used to assess attitude and values need close study, evaluation and revision. Only two districts said that there is no value.

Suggestion of Program

Over 50 districts presented programs, approaches or techniques used in their own districts which might account for any high scores they had. These districts had their assessment data above the 75th percentile or above the prediction band. Twenty-two districts felt that their high scores resulted from dedicated teachers, in-service workshops, relevant

curriculum, and various instruction such as individualized instruction, non-graded programs and flexible scheduling. Fourteen districts expressed that their high scores reflected their emphasis on basic skills and vocational education. Ten districts indicated that a combination of community, school characteristics and home background was the important factor.

Requesting Type of EQA Assistance

One hundred seven districts responded to questions about their intention of getting assistance from the EQA as a result of assessment. More than one-half of the districts asked assistance in follow-up or implementation strategies, and in developing or obtaining materials which would be helpful for enhancing goal areas. Over 22 per cent answered that assistance is adequate and said to EQA "Just continue the excellent work -- push like hell to keep the staff and to expand the program." Eighteen districts wanted further interpretation of EQA results, better presentation of materials for public relations, in-service workshop or continued advisory assistance.

Further Comments Regarding the Assessment Program

Eighty-five out of 138 districts affirmed additional comments regarding the assessment program. About one-third commended the good work of EQA saying, "Keep up the good work," "EQA has gained national recognition for its worth," and so forth. Another third recommended an improvement of instrument, expansion of resource materials, and utilization of EQA in long-range planning. Three districts expressed their opinion for more emphasis on cognitive aspects of education.

CONCLUSION

In general, as shown by the follow-up survey results, there is a strong indication that EQA results have been disseminated with an important impact on school district programs. The assessment program succeeded in providing school districts with information for a self-analysis of their educational programs and planning. Although use of EQA data is not mandated, a wide oissemination of the results have been made by a great number of districts.

Some of the significant facts to support such successful results are:

1) The assessment results have been disseminated to various persons, agencies, public relations media and other publics with many different approaches as shown on pages 3 and 4.

2) Most districts feit that EQA data has the most relevance for change in teaching strategies, course offerings and course content (page 7).

3) Most districts considered the EQA program as a means of helping them make decisions (page 8).

4) A majority of the districts requested assistance in implementation of strategies.

5) A great number of the districts commended the good work of EQA with constructive recommendations for a future plan.

6) Most districts indicated that EQA has provided a valuable tool for evaluation to measure the Ten Goals of Quality Education.

On the basis of all the data, it is concluded that the March 1974 assessment undertaken by the Division of EQA was highly successful in meeting its objectives and in benefits received by the participating districts from this evaluation opportunity.

The EQA of Pennsylvania is worthy of continued support to achieve its purpose for a development of a "whole, well-rounded individual" in the Commonwealth.

EDUCATIONAL IMPORTANCE/IMPLICATIONS OF THE STUDY

The quality of education has always been a matter of public scrutiny. People are more concerned than ever about what children are learning and how well they're learning it. Department of Education shares that concern.

Hundreds of thousands of fifth, eighth, and eleventh grade students in public schools have participated in the Department of Education's ongoing Educational Quality Assessment Program in an effort to measure strengths and weaknesses among schools throughout the state. It is important for both schools and EQA to ascertain reactions to the assessment. Thus, the study has resulted in significantly important evidences for 1) extent of dissemination of EQA results, 2) usefulness of EQA and 3) contribution of the EQA Advisory Service. Also, the study results are valuable for EQA in the process of revising procedures, measurement instruments and condition variables for post-1976 assessment.

APPENDIX

Follow-Up Opinionnaire to Superintendent

For the March 1974 Assessment

72

DISTRICT_____
POSITION_____

Educational Quality Assessment (EQA)
Follow-Up Opinionnaire to Superintendent
for the
March 1974 Assessment

1. To whom have your assessment results been disseminated? (Indicate the approximate number of people involved for each category.)

____a) school board members
____b) principals
____c) central office staff
____d) most elementary teachers
____e) most middle/junior H.S. teachers
____f) most high school teachers
____g) local service clubs (Lions, Jaycees, etc.)
____h) PTA, PTO, any parent group
____i) students
____j) general public
____k) other_____
____l) none

2. What methods have been used to inform others about the EQA report? (Mark as many as are appropriate.)

____a) special written report
____b) school district newsletter
____c) press release
____d) faculty memorandum
____e) curriculum bulletin
____f) inservice presentations
____g) school board meeting
____h) faculty meeting
____i) PTA presentation
____j) special presentation
____k) regular meeting with _____
____l) other
____m) none

3. Which of the following describe the use made of the EQA information? (Check as many as appropriate.)

____a) The information has not, as yet, been used.
____b) The information was used to reflect the favorability of our present programs.
____c) A new program is being planned for one or more of our schools as a result of the information. (List goal areas.)

____d) Revisions of some existing programs are underway as a result of the information. (List goal areas.)

____e) The information has been reviewed with building administrators for their use in program planning.

____f) The information served as a basis for teacher inservice activity. (List goal areas.)

____g) A new program has been "tried out" in one of our schools as a result of the information.
____h) A new program has been incorporated into one school's program as a result of the information.
____i) A new program has been incorporated into several of our schools as a result of the information
(If g, h or i has been checked, please list goal areas.)

4. If in item 3 you check (a), please check the statement below which best describes the situation in your district.

____a) The information was not sufficiantly credible to merit use.
____b) District personnel have not had enough time to put the information to use.
____c) The results did not contain any useful information.
____d) Other (Please describe)

5. Was the interpreting team thorough enough in its explanation of the report? ____Yes ____No ____I was not present

6. Should there have been a follow-up interpretation session?
____Yes ____No

7. Mark the five goal areas for which the information provided by the assessment was most <u>useful</u> in planning.

____Self-Esteem
____Understanding Others
____Basic Skills-Verbal
____Basic Skills-Mathematics
____Interest in School and Learning
____Citizenship
____Health
____Vocational Attitude
____Vocational Knowledge
____Creative Activities
____Appreciating Human Accomplishments
____Preparing for a Changing World

8. At which grade level(s) did you give a standardized achievement test in 1973-74 school year? (Mark as many as are appropriate)

Grade Level: 1)____ 2)____ 3)____ 4)____ 5)____ 6)____
7)____ 8)____ 9)____ 10)____ 11)____ 12)____
None)____

9. How relevant is the information provided in the report to decisions which must be made in the following areas?

	Very Relevant	Quite Relevant	Relevant	Not Relevant
a) changes in course offerings	____	____	____	____
b) changes in course content	____	____	____	____
c) change in teaching strategies	____	____	____	____
d) changes in teaching assignments	____	____	____	____
e) changes in financial allocations	____	____	____	____
f) changes in school facilities	____	____	____	____

10. The EQA information

____a) called attention to a problem area not previously noted by district staff.
____b) confirmed suspicions about district problems.
____c) did not identify any serious problems.

11. How do you consider the EQA program as a means of helping you make decisions?

____a) very useful
____b) useful
____c) not very useful
____d) useless

12. Do you believe the EQA information represents a true picture of your district in the areas assessed? ____Yes ____No

13. What do you feel is the value of EQA?

14. Can you suggest any programs, approaches or techniques used in your district that may account for any high scores you have? (Above the 75th percentile or above the prediction band)

15. Would item response frequency data be valuable if given with the initial report? ___Yes

16. Did you use criterion reference subscale scores in your explanation of the results? ___Yes ___No

17. Is condition variable data more harmful than it is helpful? ___Yes ___No If yes, which variables?

18. Were the follow-up booklets on suggested strategies of any value? ___Yes ___No ___Not received If no, which goal booklet(s) and why?

19. What type of assistance would you like from the PDE as a result of the assessment?

20. Please add any comments you wish regarding the assessment programs.

HYPOTHESIS-TESTING IN LARGE-SCALE ASSESSMENT

FRANK W. RIVAS
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Mr. Frank W. Rivas
Associate Writer
National Assessment
of Educational Progress
1860 Lincoln Street, Suite 700
Denver, Colorado 80203

HYPOTHESIS-TESTING IN LARGE-SCALE ASSESSMENT

FRANK W. RIVAS

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

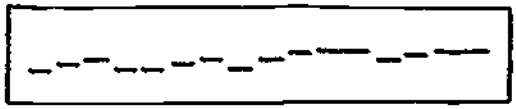
(Summary)

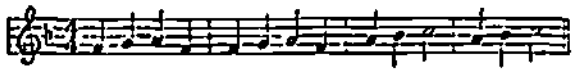
Large-scale assessment reports provide descriptive data which, even when interesting, cannot be used to improve the educational system. Often, however, assessment results are not even interesting because they merely confirm our presuppositions. These are highly subjective observations, to be sure, but observations with at least some evidence to support them. But rather than defending these statements, this paper attempts to analyze one cause for these problems and then suggests a remedy.

One reason for the plodding character of assessment reports is assessment design itself. Assessments have been designed to survey skills in broad learning areas rather than to test hypotheses about skill development. The failure to test hypotheses is not inherent in assessment methodology; hypothesis-testing can be incorporated in an assessment with relatively few modifications. Since most research in education is either theoretical or done with relatively small nonrandom samples, assessments provide the ideal vehicle for generalizing the results to a large population. Assessments can be used to test hypotheses already tested in more limited situations, thereby increasing external validity.

Hypothesis-testing is not entirely new to assessment, but the process has thus far been confined to a few isolated instances. In the national assessment of music, for example, an attempt was made to understand problems children might have

with reading music notation. Music educators hypothesized that line notation, which used height for pitch and length for duration, would be easier to understand than the more complex conventions of standard notation. To test this, the song "Are You Sleeping?" was represented in both notations as follows:





In fact, students had no more difficulty with the standard notation than with the line notation; so the hypothesis was not supported.

The national assessment of writing provides an example of using background variables for hypothesis-testing. To determine which variable most affects a student's ability to write well, the background form included questions about the number of papers the student is required to write, the type and duration of instruction on writing, whether the student typically rewrites papers before turning them in and whether returned papers include suggestions on how to improve writing. Although the data have not yet been analyzed, the results will be interesting since each of the variables represents what a number of experts hypothesize to be an important determinant of writing skills.

As these examples illustrate, hypothesis-testing can yield results both more interesting and, for some audiences, more useful than can broad surveys of skills. There are, of course, measurement problems in hypothesis-testing (as there are in surveys), but large-scale assessments generally employ measurement specialists of more than adequate skill for such designs. So the real problem is hypothesis-generation.

There are three techniques of hypothesis-generation available: (a) systematically reviewing the literature and interviewing experts in the field, (b) critically examining previous assessment results and (c) conducting qualitative (observational) field research prior to the quantitative assessment. Since the first two techniques are probably familiar to assessment personnel, the remainder of this paper will concentrate on using qualitative field research to generate hypotheses for large-scale assessments.

Qualitative research, characterized by firsthand involvement with the social world, allows one to generate hypotheses at least somewhat independent of contemporary theoretical models. Basically, there are two modes of observation: watching what people do and asking them about their actions and observations. Both direct observations and interviews range in degree of structure from virtually unstructured to strict interview schedules and full-scale observational systems. It would be convenient if there existed a literature of

qualitative studies with hypotheses applicable to large-scale assessments, but, unfortunately, such a literature does not exist. So assessments have the additional burden of completing this observational research too.

There are, however, a limited number of qualitative studies of education that have suggested hypotheses worthy of large-scale investigation. For example, in a study of Chicago elementary schools, Hariet Tamage and Robert M. Rippey[1] were surprised to find that the predicted background variables had little effect on educational achievement. Based on their observations, they hypothesized that threat of failure and the degree of socializing experience were more reliable predictors of achievement. If hypotheses like these could be supported in a large-scale assessment, the implications for education would be great.

Threat of failure and degree of socializing experience are, as noted above, more difficult to measure than the demographic variables most commonly measured in national or state assessments, but measurement is nonetheless possible. Instruments or at least prototypes for such instruments have been compiled in publications like Measurement of Affect and Humanizing of Education.[2] A measure of threat of failure

---

[1] "Elementary School Cases" in Evaluating Educational Performance, ed., Herbert J. Walberg (Berkeley, Calif.: McCutchan, 1974).

[2] Salt Lake City: Interstate Education Resource Service Center, 1974.

could be as simple as asking elementary children whether they like to display the r schoolwork in the classroom.

Qualitative research conducted prior to assessments can provide hypotheses to be tested in the assessment proper. Combining the advantages of qualitative and large-scale quantitative studies might well provide a new direction for educational research.

A PLAN FOR UTILIZATION OF ASSESSMENT DATA BY LOCAL EDUCATION AGENCIES

JOHN A. JONES AND CHARLES D. OVIATT
MISSOURI DEPARTMENT OF ELEMENTARY AND SECONDARY EDUCATION

Dr. John A. Jones
Supervisor of Assessment
Department of Elementary
and Secondary Education
Assessment Task Force
PO Box 480
Jefferson City, Missouri 65101

Dr. Charles D. Oviatt
Director of Assessment
State Department of Education
PO Box 480
Jefferson City, Missouri 65101

A PLAN FOR UTILIZATION OF ASSESSMENT DATA BY LOCAL EDUCATION AGENCIES

JOHN A. JONES and CHARLES D. OVIATT

MISSOURI DEPARTMENT OF ELEMENTARY AND SECONDARY EDUCATION

The State Education Agency (SEA) developed a plan for educational assessment which included the following phases designed to improve the quality of education in the state.

1. Goal Development - Statewide educational goals and subgoals were written and approved by the State Board of Education.
2. Objective Development - Curriculum committees were appointed to develop educational objectives directly related to the statewide educational goals and subgoals. These objectives are broad enough to be used for curriculum planning on a K-12 basis and usually are not behavioral in nature.

   After the statewide educational objectives were written, they were submitted to interested groups in the state for revision and ranking for assessment purposes.
3. Identification of Assessment Purposes - It was decided to develop an assessment instrument that tested a wide variety of knowledge, skills, and attitudes, rather than to develop a series of narrowly defined criterion-referenced instruments. The decision was made not to create individual student scores and to use assessment information only for general program planning purposes at both the SEA and LEA levels.
4. Population to be Assessed - It was decided to develop an assessment instrument for the grade 12 level and for the grade six level.

5. Instrumentation - The rankings were used in selecting about half of the educational objectives to be used as the basis for the assessment instrument. A commercial test development firm was contracted to develop performance indicators and related test items for the selected educational objectives.
6. Administration - The assessment instrument is administered by LEA personnel who have been trained by SEA staff. When statewide assessment data are collected, groups of test items were randomly assigned to students and administered in a single one-hour testing session.
7. Scoring - All of the assessment test items are multiple-choice type items and responses are made on machine-scorable answer sheets.
8. Reporting and Utilization - A School Assessment Data Summary was created for each school involved in the assessment effort for use by LEA personnel. Schools may become a part of the SEA assessment program by being selected as a part of a statewide random sample of schools or by volunteering to conduct a local assessment using the statewide assessment instrument.

The School Assessment Data Summary is composed of two parts, the Summary of Data by Subgoal and the Summary of Data by Item. The Summary of Data by Subgoal reports the frequency of correct and incorrect responses and their corresponding percentages for each subgoal. The percentage of correct responses (P-value) achieved by a statewide random sample of students is also reported for each subgoal. The Summary of Data by Item reports for each item a description of the item, school frequencies of correct and incorrect responses, school item P-value, and the statewide item P-value.

A Guide for Interpretation and Utilization of Assessment Data is sent to each school along with their School Assessment Data Summary. In this guide, a model for curriculum review using assessment information is presented, and a model for curriculum management based on educational objectives is described. The attached flowchart shows the logic of the model for curriculum review. The basic steps of this model are as follows:

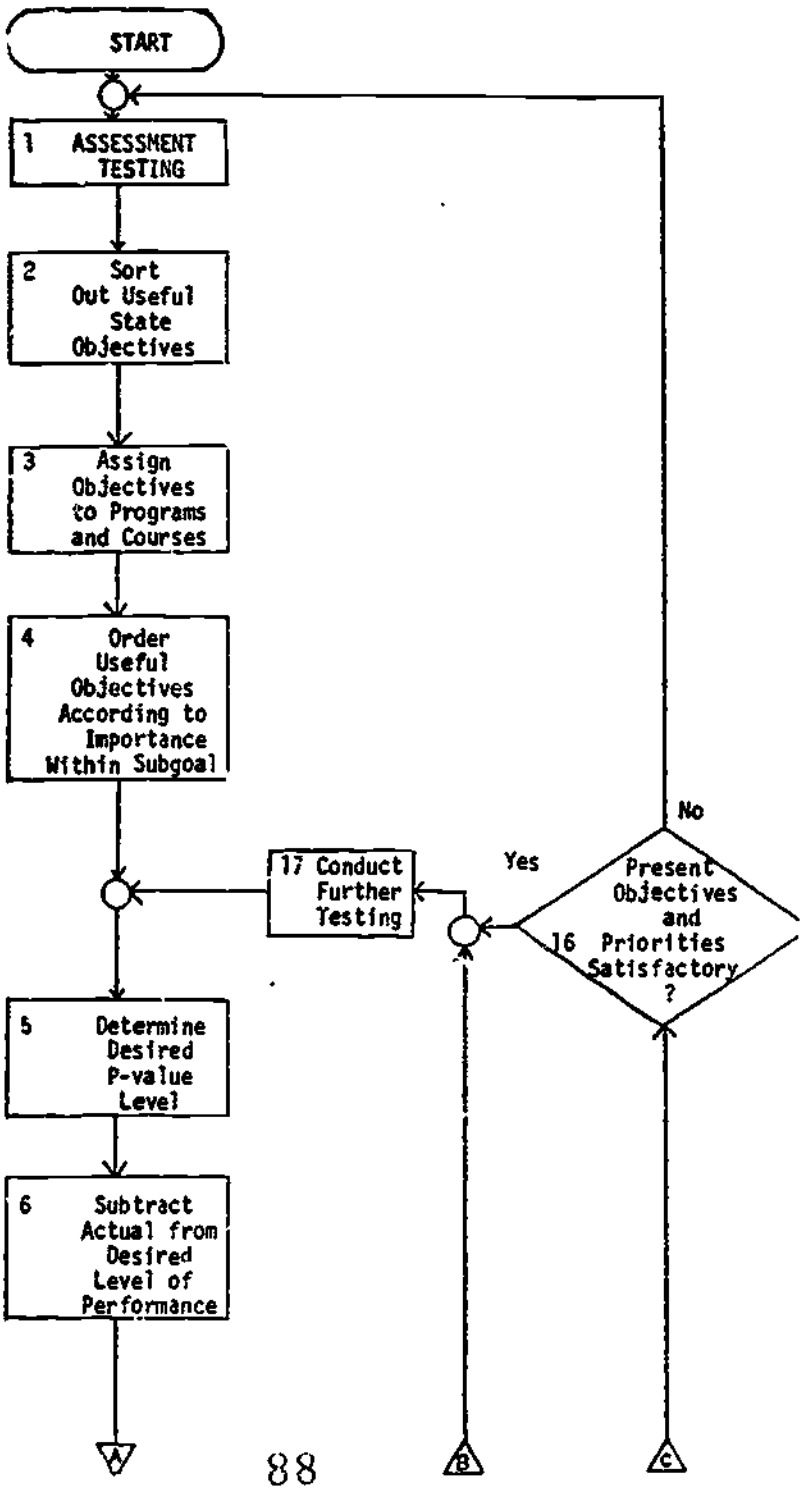
- The LEA decides to administer all or a part of the statewide assessment instrument.
- After the instrument has been administered, the LEA staff and other interested citizens identify which of the statewide educational objectives should be the responsibility of their school.
- Those educational objectives that are judged to be the responsibility of the school are assigned to the programs and courses of the school by the school staff. The school staff is divided into committees to interpret the part of the school report for which they are responsible.
- The LEA curriculum committees then rank the objectives for which they are responsible according to their curricular importance.
- The LEA curriculum committees then take the part of the school assessment data summary which gives information concerning the objectives for which the committee is responsible. These committees then rank these items according to the size of their P-values and according to the size of the differences between

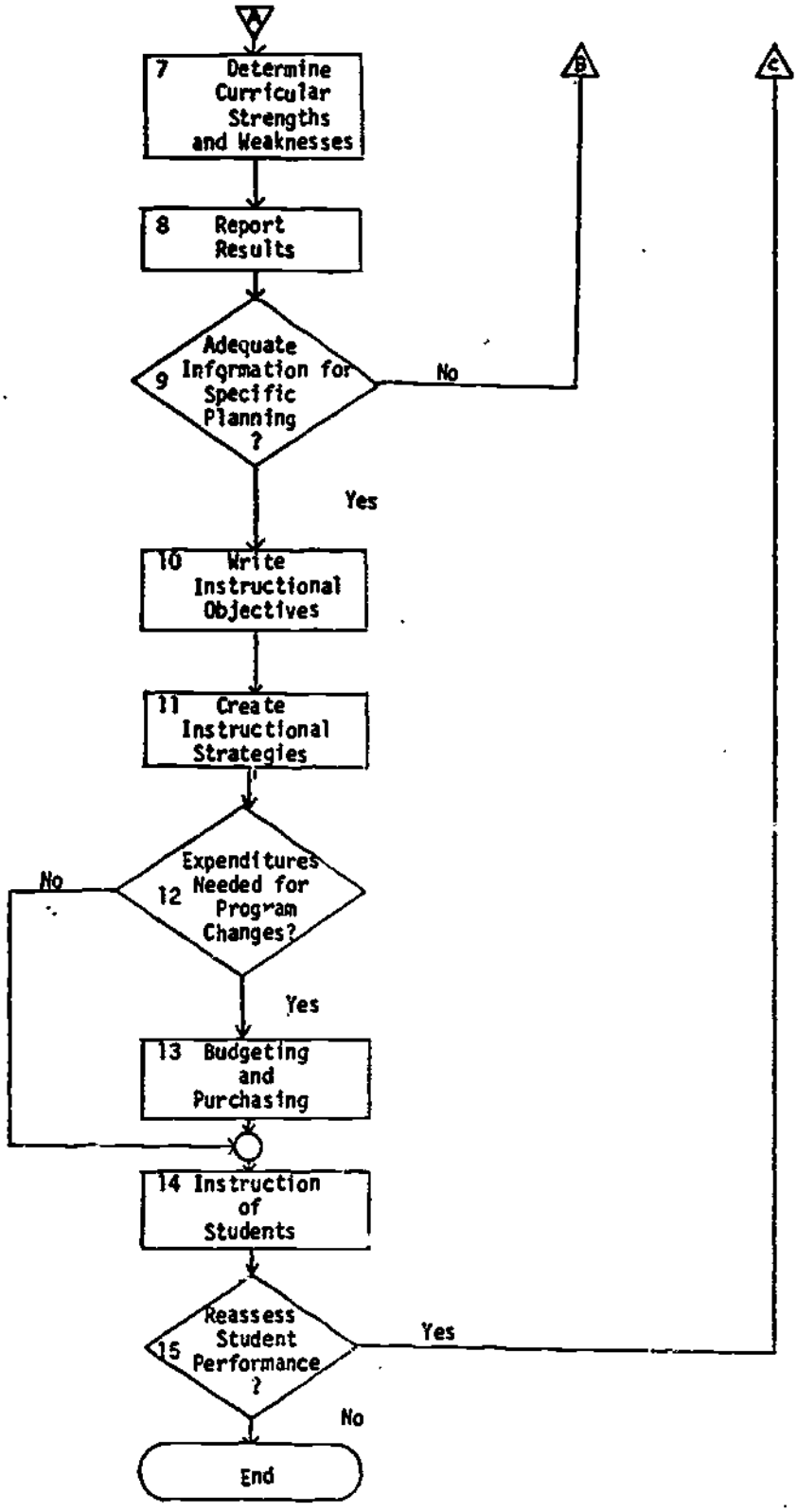
state and local item P-values. Items that consistently receive high or low rankings are used to identify relative strengths and weaknesses of students' performances. These committees then write summaries of their findings and submit them to the school administration.

- These committee reports are compiled and presented to the LEA administration and school board for their review and action. After the processes of interpretation and administrative review have occurred, the LEA administration is in a position to release the assessment results to local news media.
- If a mandate from the LEA administration is given to the school staff, specific program planning may then proceed in revising the programs of the school to correct the identified weaknesses in student performance.
- Specific program revision may identify needs for new instructional equipment and materials which require budgetary expenditures. If the LEA cannot alter its instructional expenditures, this should be announced to the school staff near the beginning of this planning process.
- Teachers should instruct students in reference to the domains of knowledge described by the educational objectives that represent identified weaknesses in student performance.
- After the LEA staff has been involved in creating plans to correct identified programmatic weaknesses, long-term leadership should be provided to help in increasing the levels of student performance in the areas of identified curricular weakness.

- Periodic reassessment of student performance is needed to give LEA personnel information to make evaluative comparisons concerning the school's progress in correcting identified weaknesses.

LOGIC MODEL FOR UTILIZATION OF ASSESSMENT INFORMATION




A
7 Determine Curricular Strengths and Weaknesses
8 Report Results
9 Adequate Information for Specific Planning ?
No
B
C
Yes
10 Write Instructional Objectives
11 Create Instructional Strategies
12 Expenditures Needed for Program Changes?
No
Yes
13 Budgeting and Purchasing
14 Instruction of Students
15 Reassess Student Performance ?
Yes
No
End

ACT TEST DATA AND PROGRAM ASSESSMENT FOR LARGE SCHOOL DISTRICTS

ROBERT CRAMER
SHAWNEE MISSION USD #152, KANSAS

Mr. R. H. Cramer
Director, Program Evaluation
U.S.D. #512
7345 Lowell
Shawnee Mission, Kansas 66204

ACT TEST DATA AND PROGRAM ASSESSMENT FOR LARGE SCHOOL DISTRICTS

ROBERT CRAMER

SHAWNEE MISSION USD #152, KANSAS

U. S. SCHOOLS A SCANDAL, TEST SCORES PLUMMET....are recent Kansas City Star headlines to articles concerned with the decrease in academic ability of the recent crop of high school seniors as measured by the highly publicized Scholastic Aptitude Testing Program. The natural question for one to ask is, "Well, how about our schools?" This report has been prepared to address such a question by providing assessment information in proper perspective or with a comparable benchmark.

For the past three years, the Department of Research and Evaluation has been conducting studies in cooperation with the American College Testing Program. (ACT). ACT is the "Avis" of the testing industry, but the program most commonly used by our students and those of comparable districts for college placement testing. This project has provided us with a far better benchmark of our graduating seniors' learning and achievement than anything we have had in the past. This benchmark study has been made possible through the selection of other schools and school districts (from Maryland to California) similar to ours in many important ways, namely, those with students of the same kinds as ours -- high income, suburban districts.

The spin-off of this project is particularly relevant in today's world of accountability and declining test scores. The aim of such studies is to bring the leadership of the educational system a little closer to an operations management philosophy by providing information for data based decision making.

In the preparation of this report, state and national results are not reported for comparison purposes due to the fact that the test scores of both this district and the "selected" group of schools are far enough above the national and state sample to make them irrelevant benchmarks. It should also be pointed out that the data for this analysis came from approximately two-thirds of the graduating class each year.

This report has been prepared in four sections, each of which provides a straightforward, factual attempt to answer the following questions:

Section 1. How well are our students learning?

Section 2. How do they feel about their schools?

Section 3. What kinds of help are they asking for?

Section 4. What are their career aspirations?

Section I

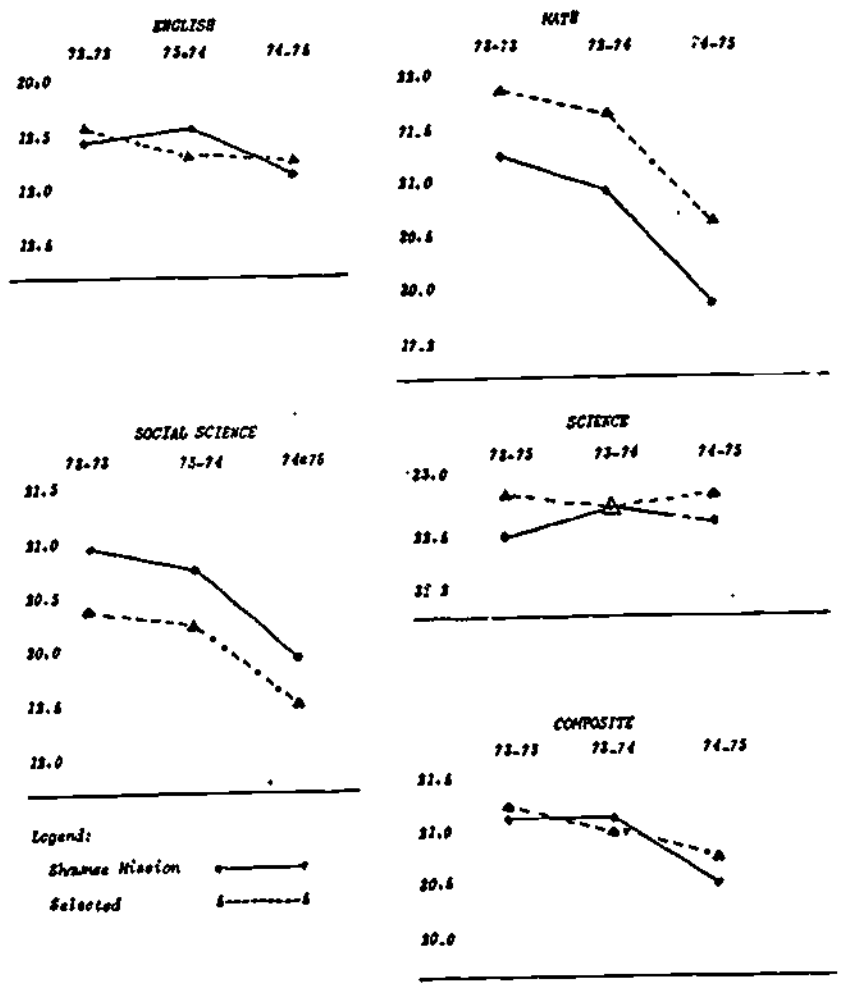
Achievement

Achievement data from both the Shawnee Mission graduates and those of the selected school districts have been summarized as Mean scores for analysis. These Mean ACT scores have been reported in the areas of English, Math, Social Science and Science, as well as a Composite. In addition to a tabular comparison of the scores between Shawnee Mission and selected schools, the scores of both groups have been displayed graphically for trend analysis.

MEAN ACT SCORES IN TABULAR FORM

	English		Math		Social Studies		Science		Composite	
	S.M.	Sel.	S.M.	Sel.	S.M.	Sel.	S.M.	Sel.	S.M.	Sel.
1972-73	19.5	19.6	21.3	21.9	21.0	20.4	22.6	22.9	21.2	21.3
1973-74	19.6	19.4	21.0	21.7	20.8	20.3	22.8	22.8	21.2	21.1
1974-75	19.2	19.3	19.9	20.7	20.0	19.6	22.7	22.9	20.6	20.6

MEAN ACT SCORES IN GRAPHICAL FORM



Interpretation:

English scores are consistent with national trends showing a very slight downward trend across three years for both groups. However, the Shawnee Mission student performs equally as well as students of other top schools.

A steady decline in Math scores is apparent in both groups. Further, our students show a lower level of achievement over the three years as noted by the parallel slopes of the two lines.

In Social Studies, both groups show a steady decline across years. However, Shawnee Mission students show achievement above the selected group across all three years.

Natural Science scores are contrary to the national picture in that they remain nearly constant over the three years. There does not appear to be any significant difference in Science achievement between Shawnee Mission students and those of the selected schools.

The overall composite scores show the same general downward trend which is consistent with national trends. However, the graph shows that our students are virtually the same in achievement with their suburban counterparts.

Summary:

In summary, the general downward trend in achievement as depicted nationally is also apparent from these results with the exception of Natural Science and to a lesser degree in English. When compared to the selected group, Shawnee Mission appears to be:

(1) STRONGER in Social Studies,
(2) Same in Natural Science,
(3) Same in English,
(4) WEAKER in Math.

The information suggests that Shawnee Mission maintains at least the same degree of excellence in Natural Science, Social Science, and English programs as do other top schools across the country. In Mathematics, the performance of our students exceeds both national and state averages, but when compared with comparable school districts, there is an indication of need for improvements. This fact is also borne out by results of the Iowa Tests of Basic Skills at lower levels where both Map Skills and Math Skills appear to be weaker among the skills tested.

To provide an example of some of the skills tested by the ACT, sample test items for each area are attached.

Section II

Student Attitudes and Impressions Concerning Their School

For the past two years, ACT has provided valuable feedback information regarding how students feel about their school. This information has been summarized from data collected from the students tested including their responses to questions of the expressed adequacy of their high school education according to their high school curriculum or program; and their degree of satisfaction with various aspects of their high school.

The following two tables summarize the results of student responses as to the adequacy and their satisfaction with the high school program.

Summary:

Responses of both groups indicated that their high school education was slightly less adequate in 1974-75 than in 1973-74. Shawnee Mission students, however, considered their high school education to be no more or less adequate than their counterparts across the country.

Curiously, both groups seem to be more satisfied with various aspects of their high school in 1974-75 than in 1973-74, even though they expressed that their high school education to be slightly less adequate. By and large, over the past two years Shawnee Mission students have similar feelings about their school as do other suburban students across the United States.

Expressed Adequacy of HS Education According to HS Curriculum or Program

% Student Responses

Expressed Adequacy		Bus-Comm 73-74	Bus-Comm 74-75	Voc-Occup 73-74	Voc-Occup 74-75	Coll Prep 73-74	Coll Prep 74-75	General or Other 73-74	General or Other 74-75	Avg ACT Comp 73-74	Avg ACT Comp 74-75
EXCELLENT	Shown this	13	9	13	9	25	22	8	8	24	23
	Selected	20	14	11	11	31	26	11	12	24	23
GOOD	Shown this	53	50	52	45	57	51	51	49	22	21
	Selected	48	43	52	41	53	47	48	40	22	21
AVERAGE	Shown this	23	23	28	24	15	19	33	29	19	19
	Selected	29	25	30	28	14	17	35	32	18	18
BELOW AVERAGE	Shown this	10	4	6	11	2	3	6	6	18	17
	Selected	6	6	6	8	1	2	4	8	18	18
VERY INADEQUATE	Shown this	1	14	1	10	1	11	2	7	19	21
	Selected	2	12	1	9	1	10	1	8	20	21

Student Satisfaction with Various Aspects of the Local HS

(% Student Responses)

		SATISFIED, NO CHANGE NECESSARY		PRETTY MUCH NEUTRAL		DISSATISFIED, IMPROVEMENT NEEDED		NO EXPERIENCE	
		73-74 total %	74-75	73-74 total %	74-75	73-74 total %	74-75	73-74 total %	74-75
Classroom Instruction	Stu Mem.	52	56	27	27	20	16	1	1
	School	51	54	27	25	22	18	0	1
No. & Variety of Course Offerings	S.-M.	69	71	14	14	17	14	C	1
	School	73	77	12	11	15	12	C	1
Grading Practices & Policies	S.-M.	41	48	28	26	31	35	C	1
	School	35	41	26	25	39	30	1	1
No. & Kinds of Tests Given	S.-M.	46	48	36	35	19	17	0	1
	School	46	44	36	35	24	20	C	1
Guidance Services	S.-M.	35	37	24	21	34	37	5	4
	School	46	46	24	23	26	26	5	5
School Rules, Regulations, & Policies	S.-M.	129	29	20	22	50	45	1	1
	School	38	35	23	24	46	40	2	1
Library or Learning Center	S.-M.	42	47	23	23	33	27	2	2
	School	62	62	21	20	19	16	1	2
Laboratory Facilities	S.-M.	46	50	27	26	11	12	19	13
	School	57	57	24	23	11	11	8	8
Provisions for Spec Help in Reading, Math, etc.	S.-M.	36	33	24	20	20	18	27	29
	School	32	33	24	22	18	17	26	27
Provisions for Acad Outst Stu	S.-M.	43	43	24	21	12	12	21	24
	School	53	53	23	20	7	7	17	20
Adequacy of Prog in Educ & Planning	S.-M.	30	33	37	29	35	26	8	11
	School	40	42	37	26	28	23	5	10

Section III

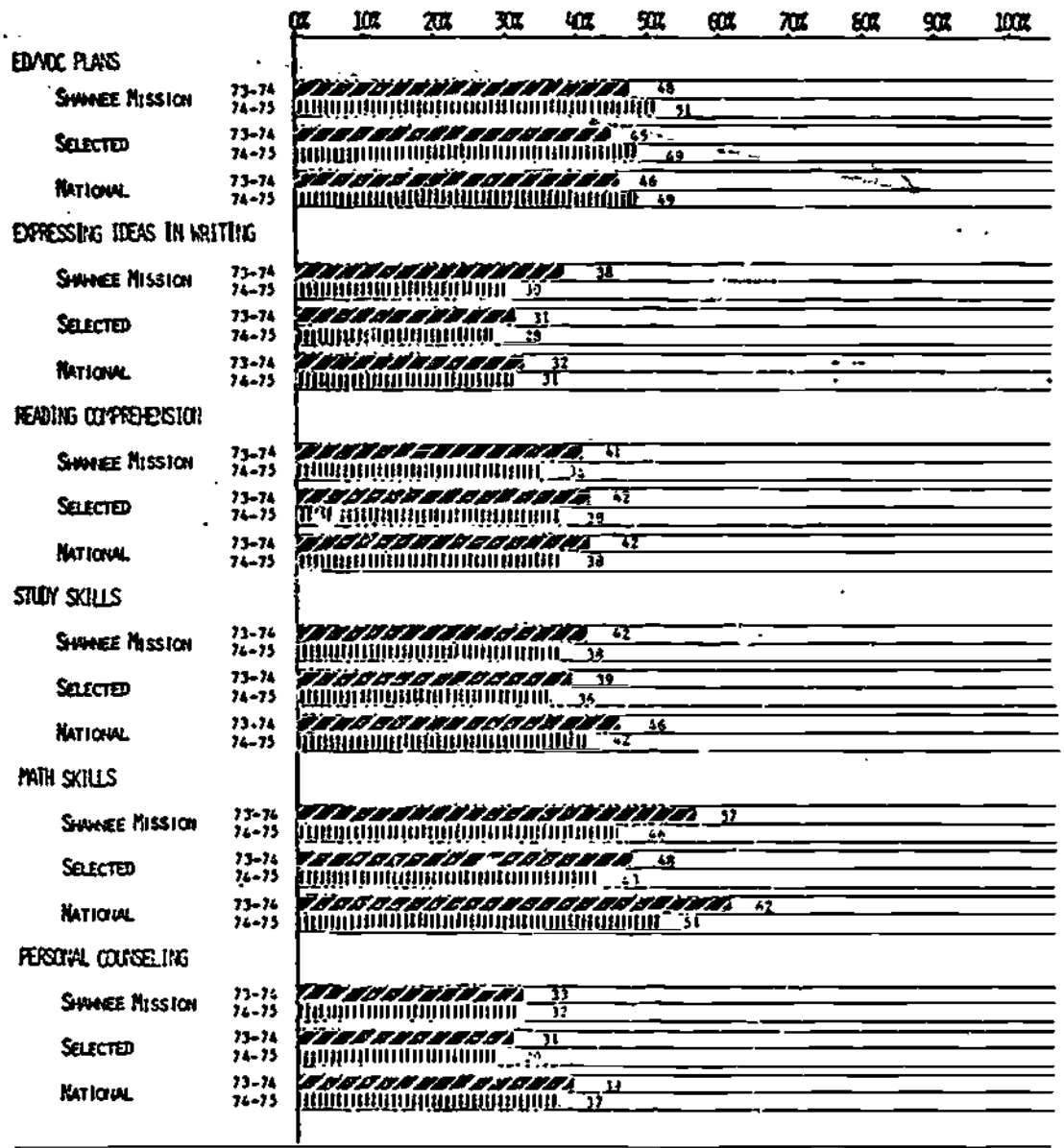
Requests for Educational Assistance

One task of our schools is to help individuals cope with their own educational problems . . . . . especially those involving aspects which are prime responsibilities of schools and for which patrons hold them accountable. The following chart shows the percent of students who requested educational assistance with each of the selected program areas.

Summary:

The greatest number of students in all groups are asking for assistance with their educational/vocational pl. s, i.e., career education. Next in line as a priority need for assistance are Math Skills. These needs are also consistent with the achievement results as well as the expression of student attitudes and impressions of their school.

PERCENT OF STUDENTS WHO REQUESTED EDUCATIONAL ASSISTANCE



Section IV

Career Aspirations

This section summarizes the responses of students regarding their career aspirations. Table A shows the percentage of students indicating a proposed education major in selected career clusters. Table B shows the degree aspirations of the groups. Table C shows the degree of confidence the students have in their decisions regarding their planned education major and their first career choice. Table D shows the average number of out-of-class activities for the ACT tested students.

Summary:

In reviewing the results of the career aspirations of the students tested, it was noted that there were close parallels between Shawnee Mission students' proposed education majors and those of the selected schools. Both groups departed from the national results only in the health professions and social science.

In regard to educational degree aspirations, again there were close parallels between the two groups. These indicators confirm the validity of the benchmark as being useful for comparison purposes.

In summarizing the results of the confidence students placed in their planned educational major and first vocational choice, men showed fairly consistent results indicating agreement in career aspirations and the educational means to attain their career choice. For women, the 1974-75 groups seem to be more certain about their plans for education and career choices than the 1973-74 groups. Overall, our students are very similar to the selected group in educational plans and career aspirations. This tells us that our counseling efforts are as effective as the benchmark group.

The summary of the average number of out-of-class activities indicates what constitutes an important dimension of talent among the students but appears to be relatively uncorrelated with achievement scores in English, Math, Social Studies and Natural Science. Nonetheless -- they represent a component of our students' behavior and interest of extreme importance today. Inspection of the data revealed an interesting profile when reflected against the benchmark students. Our men students responded as taking part less in athletics and more in leadership, music, science, work experiences and curiously, writing. Our women students take part significantly more in work experiences; in no case do they participate less. A special note was made of the dramatic increase in athletic participation by 1974-75 women students over 1973-74, and more in line with the selected group index. Compared nationally, both groups show a significant difference. As expected, our students and the students from the selected schools are much alike but depart from the national profile. Nationally, men students in 1974-75 were much more involved in athletics, community service, leadership, and speech, while slightly less were involved in work experiences. Nationally, women students in 1974-75 were more involved in community service, leadership, and music. Other categories were nearly the same.

One conclusion that could be drawn is that students in the suburban schools have to expend more time on their academic studies. They buy this time by participating less in athletics if they are a man, and less in community service, leadership, and music if they are a woman.

Table A

Proposed Education Major

Career Cluster	Shawnee Mission		"Selected"	
	73-74	74-75	73-74	74-75
Agriculture/Forestry	4%	4%	4%	4%
Architecture	2	2	3	3
Biological Science	4	4	5	4
Business and Commerce	16	13	16	16
Communications	3	4	3	3
Computer & Info. Science	1	1	1	1
Education	13	10	11	10
Engineering	6	6	5	5
Fine Arts & Applied Arts	8	9	8	7
Foreign Languages	1	1	1	1
Health Professions	13	13	14	14
Home Economics	3	3	2	2
Letters (Humanities)	2	2	1	1
Physical Science (Physics, Chemistry, Geology)	2	1	2	2
Mathematics	1	1	1	1
Community Service	2	3	3	3
Social Sciences	11	11	11	10
Trade, Indust., Technology	2	2	1	2
General Studies	-	2	-	3
Undecided	6	7	7	8

Table B

Educational Degree Aspirations

Level	Shawnee Mission 73-74	Shawnee Mission 74-75	"Selected" 73-74	"Selected" 74-75
*Vocational or Technical Program (Less than two years)	1%	1%	1%	1%
*Two-Year College Degree	9	6	8	7
Bachelor's Degree	44	47	41	44
*One- or Two-Year Graduate Study (M.A., MBS, etc.)	20	22	22	23
Professional Level Degree (PH.D., M.D., LL.D., J.D., etc.)	21	21	22	21
Other	5	3	6	4

*The levels where these groups deviate from the National profile.

Table C

Students' Confidence in Their Planned Educational Major and First Career Choice

C-1 Men

	Shawnee Mission				"Selected"			
	Planned Ed. Major		First Career Choice		Planned Ed. Major		First Career Choice	
	73-74	74-75	73-74	74-75	73-74	74-75	73-74	74-75
Very Sure	31%	28%	25%	27%	--	25%	--	26%
Fairly Sure	46	48	47	46	--	49	--	46
Not Sure	23	22	27	37	--	24	--	38

--not broken out by sexes in 73-74, see total.

C-2 Women

	Shawnee Mission				"Selected"			
	Planned Ed. Major		First Career Choice		Planned Ed. Major		First Career Choice	
	73-74	74-75	73-74	74-75	73-74	74-75	73-74	74-75
Very Sure	35%	34%	29%	35%	--	32%	--	38%
Fairly Sure	46	47	47	48	--	47	--	46
Not Sure	19	21	27	26	--	21	--	27

C-3 TOTAL (Men and Women)

	Shawnee Mission				"Selected"			
	Planned Ed. Major		First Career Choice		Planned Ed. Major		First Career Choice	
	73-74	74-75	73-74	74-75	73-74	74-75	73-74	74-75
Very Sure	33%	31%	27%	28%	29%	28%	25%	27%
Fairly Sure	46	46	47	48	49	47	47	47
Not Sure	21	23	26	26	22	23	28	23

Table D

Average Number of Out-of-Class Activities

Accomplishment	Men				Women				Total			
Area	S.N.		Selected		S.N.		Selected		S.N.		Selected	
	73-74	74-75	73-74	74-75	73-74	74-75	73-74	74-75	73-74	74-75	73-74	74-75
Art	.69	.71	.83	.75	1.02	.92	1.18	1.12	.87	.83	1.01	.95
Athletics	3.00	2.84	3.11	2.99	1.83	2.10	2.01	2.03	2.37	2.44	2.55	2.48
Community Service	.81	.85	.82	.90	1.34	1.40	1.52	1.41	1.10	1.16	1.16	1.17
Leadership	1.45	1.19	1.25	1.10	1.47	1.37	1.46	1.22	1.46	1.29	1.36	1.16
Music	1.28	1.22	1.09	1.15	1.58	1.64	1.55	1.56	1.43	1.45	1.33	1.39
Science	.57	.52	.45	.46	.32	.39	.31	.30	.44	.45	.38	.37
Speech	.63	.59	.51	.49	.80	.62	.73	.64	.72	.60	.63	.57
Work Experience	2.62	2.55	2.35	2.46	1.98	2.01	1.84	1.88	2.28	2.25	2.09	2.15
Writing	.76	.75	.66	.65	.96	1.06	1.05	1.00	.87	.92	.86	.83

TYPICAL SCIENCE TEST QUESTIONS

A series of experiments was designed to determine how bats are able to fly at night without colliding with obstacles. Bats were released in a closed room across which were strung fine wires adapted to register every time they were touched by one of the bats. The bats were released in the room under the following conditions.

Experiment 1
The room was well illuminated.

Experiment 2
The room was completely darkened.

Experiment 3
The room was darkened and the bats' eyes were sealed with soft black wax.

Experiment 4
The room was darkened, the bats' eyes were waxed closed, and numerous small radar transmitters were in operation throughout the room.

Experiment 5
The radar transmitters were replaced with loudspeakers which emitted high-frequency sound waves. The room was dark and the bats' eyes were waxed closed.

Experiment 6
The lights were turned on and the bats, without wax on their eyes, were released while the loudspeakers were still producing high-frequency sounds.

On the basis of these experiments, the following observations were made:

In experiments 1 through 4, the bats did not collide with the wires.

In experiment 5 the bats seemed confused and frequently collided with the wires.

In experiment 6 the bats were initially confused and collided with the wires, however, the number of collisions soon decreased.

6. Which of the following conclusions, if any, can be drawn from experiment 1?

F. Bats need light to see where they are going.
G. Bats need sound waves in order to avoid obstacles.
H. Bats can see in the dark.
J. None of these

7. Which of the following conclusions can be drawn from experiment 3?

A. Bats evidently use some sort of radar to guide themselves.
B. The presence of radar waves has no apparent effect on the bats.
C. The presence of radar waves confuses the bats by obstructing their natural means of locating obstacles.
D. None of these

8. Which experiment or group of experiments listed below shows that bats can ordinarily fly safely without using their eyes?

F. 1 only
G. 1 and 2
H. 1 and 3
J. 1, 2, and 3

9. Which of the following is true regarding the statement: *bats are not normally aware that daylight interferes with their ability to avoid obstacles*?

A. The statement agrees with the data.
B. The statement is contradicted by the data.
C. The statement cannot be judged without more data.
D. The statement is an experimental assumption.

The following questions are not based on a reading passage. You are to answer these questions on the basis of your background in the natural sciences.

10. The emergence of new strains of houseflies capable of withstanding the poisonous effects of the chemical DDT is an example of:

F. adaptation.
G. the Mendelian law.
H. implementation.
J. regeneration.

11. What is the main difference between a gas and a liquid?

A. Molecular weight
B. Shape of the particles
C. Geometric arrangement of the molecules
D. Average distance between the molecules

12. How were the coral reefs of tropical seas formed?

F. By the accumulation of the remains of small marine animals
G. By the erosion of islands by wind and sea
H. By the accumulation of salts and minerals precipitated by the sea
J. By undersea earthquakes

13. A warm breeze may seem cool to a bather who has just come from the water because:

A. water is a good conductor of heat.
B. moisture from the sea condenses on the skin and cools it.
C. the evaporation of water from the wet skin absorbs heat.
D. water is denser than air.

TYPICAL ENGLISH TEST QUESTIONS

DIRECTIONS:

Questions on the English Usage test are based on passages which contain technical errors and inappropriate expressions; you are to decide how they can be corrected or improved. The passages are presented in a spread-out format in which various words, phrases, and punctuation have been underlined and numbered. In the right-hand column, opposite each underlined portion, you will find a set of responses whose number corresponds to that of the underlined portion. Each set of responses contains a "NO CHANGE" option and three alternatives to the underlined version. Since your judgment about the correctness or appropriateness of a response will sometimes depend on your reading several of the sentences surrounding the underlined portion, first read through the entire passage quickly to determine its content. Then reread the passage slowly and carefully. As you come to each underlined portion during your second reading, look at the alternatives in the right-hand column and decide which of the four words or phrasings is best for the given context. Make sure you have read ahead far enough to make an accurate choice. If you think that the original version (the one in the passage) is best, blacken the oval marked A or F in the corresponding row on the test answer sheet. If you think that an alternative version is best, blacken the oval whose letter corresponds to the alternative that you have chosen as best. In every case, consider only the underlined words, phrases, and punctuation marks; you can assume that the rest of the passage is correct as written.

PASSAGE 1

Thor Heyerdahl became famous for a unique sailing expedition, which he later described in *Kon-Tiki*. Having developed a theory that the original Polynesians had sailed or drifted to the South Sea Islands from South America, it then had to be tested. After careful study he [1] built a raft that was as authentic as possible. Using only primitive equipment, he and five other men sailed into the South Seas from Peru, which he judged to be in the same [2] general area as the land of the original Polynesians. As a result, his group and him will [3] may be remembered not only as thorough scientists but also as courageous men.

Heyerdahl's courage was first tested in Ecuador. His search for trees that was large enough for the expeditionary [4] raft sent him to Quito, a city high in the Andes. There, he and his companions were warned about headhunters and bandits on the trail. Feeling undaunted, they hired a driver [5] and jeep from the U.S. Embassy, going on with their dangerous task.

After the raft was done, Heyerdahl made [7] final preparations for the expedition. Even before his crew came aboard, the courage which Heyerdahl possessed was [8] tested again. As the raft was being towed out of the harbor, it drifted under the stern of a tug. Heyerdahl had to struggle to save it.

Dangers at sea were present, but Heyerdahl [9] and his men did not show fear. Instead they developed games that were actually tests of courage. Although man-eating fish were nearby, the men swam to relieve their tension, maintaining that the fish were not [10] dangerous unless a man had already been cut or scratched. One game consisted of luring sharks within reach, catching them, and then they would yank it onto the raft. [11]

1 A. NO CHANGE
B. he set out to test it.
C. it was decided that it must be tested.
D. the theory was then to be tested.

2 F. NO CHANGE
G. Peru, being judged as
H. Peru, which had been
J. Peru judged as being

3 A. NO CHANGE
B. him and his group
C. his group and himself
D. he and his group

4 F. NO CHANGE
G. which would be of sufficient size
H. of adequate size
J. of certainly sufficient size

5 A. NO CHANGE
B. trail. Undaunted, they
C. trail, but they were undaunted, and
D. trail; undaunted they

6 F. NO CHANGE
G. Embassy; and went on with
H. Embassy and proceeded with
J. Embassy, and kept on

7 A. NO CHANGE
B. When the raft was ready,
C. The raft was speedily completed and
D. The raft having been constructed,

8 F. NO CHANGE
G. Heyerdahls' manly courage
H. Heyerdahl's courage
J. the courage of this man

9 A. NO CHANGE
B. (Begin new paragraph) Dangers at sea
C. (Do not begin new paragraph) At sea, dangers
D. (Begin new paragraph) At sea, dangers

10 F. NO CHANGE
G. tension. Maintaining
H. tension. He maintained
J. tension, because it was maintained

11 A. NO CHANGE
B. then to yank it
C. and then to yank them up
D. and yanking them

TYPICAL SOCIAL STUDIES QUESTIONS

TEST 5: Social Studies Reading

Directions:

The Social Studies Reading test samples your ability to comprehend, analyze, and evaluate reading materials in such social studies fields as history, political science, economics, sociology, anthropology, and archeology. To answer these questions, you have to draw on your background in social studies as well as on your ability to understand new material. In addition to the questions based on reading passages, there are some questions that test your general background knowledge in social studies.

Read the passage through once. Then turn to it as often as necessary to answer the questions.

Over the past several decades the growth of the U.S. economy has been marked by expansion of metropolitan areas and by "regionalization" of production—that is, a more even geographical distribution of industries over the United States. Such rapid growth causes drastic changes in the geographical structure of metropolitan areas. Manufacturing industries, which were initially attracted to the core of the city by the proximity of the railroads, a steady labor supply, and the economic advantages of mass production, are now moving toward peripheral locations.

No single explanation can be given for this trend toward suburbanization, but as cities have grown, the supply of undeveloped land has decreased. The advantages of the central metropolis continue to attract economic activity, but congestion in the central city and the development of production techniques which demand more space have tended to push industry into the suburbs. The net result has been a pattern of geographical specialization within metropolitan regions. The central city increasingly becomes zoned to white-collar and service activities, and the periphery attracts manufacturing, transportation, and other blue-collar job activities.

The development of residential areas has followed industrial movement to some extent, but suburban living (undoubtedly desired for its amenities) is still largely reserved for those who can afford it. Consequently, the central city has been losing middle- and upper-income families to the suburbs. Now people can live in dispersed residential locations: rising incomes and the proliferation of automobiles have made this both economically and technically feasible. However, this "urban sprawl" creates serious financial problems. Since tax-paying industry has fled to the suburbs, the central city has had to bear the cost of public assistance payments and other welfare service for low-income groups.

When housing developers began building on a large scale, many suburbs rapidly doubled and tripled in size. This new population required more schools and teachers, more fire and police protection, and sizable expenditures for water and sewer lines and roads. Frequently, these towns were entirely dependent on property taxes for their revenues.

To meet ever increasing expenses and broaden their tax base, some communities have tried to attract new industry. However, when town officials found themselves competing intensely for these industries, they often conceded partial exemption from property taxes to new industry in order to bargain more favorably. As a result, an area often found its tax base weakened rather than strengthened by wooing new industry. As a consequence of all these changes, both the suburbs and the central city are entangled in thorny financial problems.

1. According to the author, a rise in wages earned by employees of service industries will principally tend to:
   A. increase the physical separation between zones of residence and zones of work.
   B. decrease the tax revenues of the suburbs.
   C. decrease the tax revenues of the metropolitan areas.
   D. increase the work force in the periphery.

2. The most efficient way to solve the financial problems of a metropolitan area would be to:
   F. cut personal taxes in central cities.
   G. cut personal taxes in the suburbs.
   H. decrease public expenditures in central cities.
   J. place the entire area under one fiscal authority.

3. Which of the following problems should be given first consideration on the basis of the changing urban structure outlined in the passage?
   A. Commuter traffic between areas of residence and areas of work
   B. Highway passenger traffic between two metropolitan areas
   C. Congestion due to heavy truck traffic in downtown areas
   D. The centralization of railroad freight stations in downtown areas

TYPICAL MATHEMATICS TEST QUESTIONS

DIRECTIONS:

In the Mathematics Usage test you are to solve each problem and then choose the correct answer. In some of the problems the fifth alternative, E or K, is "None of these." In those problems, if your answer is not among the first four alternatives, blacken oval E or K.

1. Two wells pump oil continuously. One produces 4000 barrels of oil per day, which is 33 1/3 percent more than the other well produces. How many barrels of oil are produced daily by the two wells?

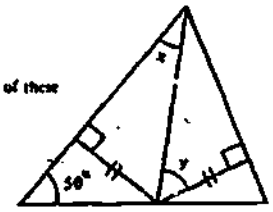
   A. 5333 1/3
   B. 6666 2/3
   C. 7000
   D. 8333 1/3
   E. 9000

2. If a car travels a miles in b minutes, how many minutes will it take to travel c miles?

   F. c/a
   G. c/b
   H. c/ab
   J. ab/c
   K. cb/a

3. In the figure below, what is the sum of the measures of the angles marked x and y?

   A. 90°
   B. 100°
   C. 130°
   D. 140°
   E. None of these



4. A man purchased 100 shares of stock at $5 a share. Each share rose 10 cents the first month, decreased 8 cents the second month, and gained 3 cents the third month. What was the value of the man's investment at the end of the third month?

   F. $505
   G. $520
   H. $525
   J. $1545
   K. None of these

5. The following multiplication scheme uses symbols other than the usual numerals. Δ corresponds to which base-10 numeral?

$$\Delta \times \theta = \theta$$
$$\theta \times \Delta = \theta$$
$$\Delta \times \Delta = \Delta$$

   A. 0
   B. 1
   C. 2
   D. 5
   E. 10

6. What is the length of a 144 degree arc in a circle whose circumference is 60 inches?

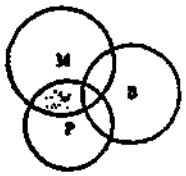
   F. 24 inches
   G. 12π inches
   H. 12π inches
   J. 36 inches
   K. 36π inches

7. What does x equal in the following equation?

$$\frac{1}{x} = \frac{1}{5} - \frac{1}{x}$$

   A. $\frac{1}{10}$
   B. $\frac{1}{5}$
   C. 5
   D. 10
   E. None of these

8. In the universe of all people, let circle M represent all Mary's friends, circle B all Bill's friends, and circle P all Pete's friends. What is represented by the shaded portion of the figure?



   F. All the people who are friends of Mary, Bill, and Pete
   G. All the people who are friends of Mary and Pete
   H. All the people who are friends of Mary and Pete but not of Bill
   J. All the people who are friends of Pete but not of Bill
   K. All the people who are not friends of Bill

9. A ship moving due north travels a course that is 12 miles from an island at its closest point. If a gun on shore has a firing range of 13 miles, for how many miles will the ship remain within range of the gun?

   A. 1
   B. 5
   C. 10
   D. 9
   E. None of these

AN EXAMPLE OF THE USE OF MULTIPLE MATRIX SAMPLING PROCEDURES IN A LOCAL DISTRICT ASSESSMENT PROGRAM

CARL D. NOVAK
LINCOLN, NEBRASKA PUBLIC SCHOOLS

Dr. Carl D. Novak
Senior Evaluator
Educational Service Unit #18
Lincoln Public Schools
720 South 22nd
Lincoln, Nebraska 68506

AN EXAMPLE OF THE USE OF MULTIPLE MATRIX SAMPLING PROCEDURES IN A LOCAL DISTRICT ASSESSMENT PROGRAM

CARL D. NOVAK

LINCOLN, NEBRASKA PUBLIC SCHOOLS

One of the critical limitations in the conduct of assessment in elementary and secondary schools is the time, expense and general disruptiveness of the data collection efforts. The demand for more and more relevant data has been an outgrowth of increased emphasis on program evaluation and accountability. The cumulative effect has been a significant increase in the amount of school time devoted to testing and data collection efforts. In many districts the point has been reached where teachers and administrators are no longer receptive to additional data collection.

The problem is complicated by the fact that many of the data collection efforts are unduly cumbersome. Administrators of assessment programs, local evaluators and directors of testing too often fail to use sampling techniques effectively. For example, program evaluation typically focuses on the performance of groups rather than the performance of individuals. Yet evaluation data is typically collected on individuals and the scores aggregated to estimate group characteristics. Similarly assessment is typically concerned with the performance of some intact group, i.e., class, school, district, or region. Yet too often assessment, particularly locally administered assessment, focuses on individuals as the data collection unit. This paper discusses an assessment effort in which data was collected on specific groups, i.e., grade levels within schools, without administrating the complete instrument to students. The procedure involved the use of multiple matrix sampling.

Objectives of the Study

The objectives of the study were to (1) demonstrate the feasibility of using multiple matrix sampling procedures to efficiently and unobtrusively collect assessment data at both the district and building levels; (2) determine whether or not data collected through the use of matrix sampling is credible to principals and teachers; and (3) test the feasibility of using such assessment data to help manage curriculum in a highly decentralized district.

Instrument

The test used was the American Association for Health, Physical Education and Recreation (AAHPER) Cooperative Health Education Test, Form 3A, published by the Educational Testing Service. The AAHPER Cooperative Health Education Test was designed as an end-of-course test to measure achievement in health education at the upper-elementary and junior high school levels. Form 3A, which was developed for use with junior high students, consists of 60 multiple choice items. The items are distributed across eleven content areas: Consumer Health, Community Health, International Health, Disease and Disorder, Personal Health Care, Sex Education, Growth and Development, Nutrition, Mental Health, Drug Use and Abuse and Safety and First Aid.

Separate National Norms are provided by sex for grades 7, 8, and 9. Normative information is also provided on each item by content area.

Procedures

The 60 item, 40 minute test was randomly divided into six 10 item matrix tests. The tests were distributed to all elementary and junior high school buildings in the district where they were administered to students in conjunction with a normal 45 minute class period. The six tests were placed in random order prior to distribution to schools.

Separate machine scorable answer sheets were printed to simplify scoring and preparation for analysis. The answer sheets were distributed after the tests had been distributed. Information was collected on school, grade level, sex, matrix test number, and item responses. Although the tests were not used to evaluate students, names were requested to insure that students took the test seriously. The entire procedure took approximately ten minutes.

The matrix tests were used to estimate (1) the school mean, standard deviation, and percentile distribution by grade level for each of the 41 participating schools and (2) the district mean, standard deviation, and percentile distribution by grade level. Since sex differences on the AAHPER Cooperative Health tests are significant and since separate norms are provided for boys and girls, the distributions for boys and girls were estimated separately. The mean, standard deviation and percentile distribution were subsequently computed for 110 groups.[1] Approximately 6,000 students (2000 6th graders, 2000 7th graders, and 2000 9th graders) participated in the Spring 1975 assessment.

---

[1] 6th grade boys and 6th grade girls in 32 attendance centers (64 groups), 7th grade boys, 9th grade boys, 7th grade girls, and 9th grade girls in 10 attendance centers (40 groups), and district boys and district girls for grades 6, 7 and 9 (6 groups).

In addition to testing the technical feasibility of using multiple matrix sampling instead of census testing, the study focused on two practical considerations, i.e. would the results be perceived as credible by administrators and building staff and would the results be useful?

Teachers and administrators within the district were accustomed to data collection in which either the entire instrument was administered to every student (the normal district testing program) or only a sample of students was administered the entire instrument. Neither of the procedures for collecting data is entirely acceptable. The district has been making a conscientious effort to reduce the amount of required student testing time to a minimum. For example, students are no longer required to respond to an entire test battery, but instead are required to take only subtests that teachers and administrators indicate that they can and will use. The use of traditional or student sampling procedures reduces the total amount of student test time but is even more disruptive than census testing. The teacher with just a few students absent for testing is faced with the dilemma of providing instruction for the remaining students and making provisions for the missing students to catch up.

The problems associated with the use of test data are not unique to assessment programs. One of the reasons the district is minimizing the present testing program is that there is little evidence that test information has been used in the past. In order to facilitate use of the information special reporting procedures were developed. Traditionally, test scores are reported by student in terms of grade equivalents, percentiles, and stanines. School stanine distributions are also made

available to local building administrators. The reporting procedures developed for the assessment data focused on group performance rather than on the performance of individuals. School means and standard deviations were reported by sex by grade level. District and national means and standard deviations were also provided to add perspective to the school scores. The percent of students scoring at or above the 25th, 50th and 75th percentile were provided using both district and national norms. Initial plans called for stanine distributions to be provided by sex by grade by school. The stanine distributions were produced, however, nine data points proved to be too many to be conveniently interpreted. The 25th, 50th and 75th percentile system represents a more workable compromise.

Results

Estimates of the mean score, the standard deviation and percentile distributions were computed by sex by grade level by school. The district mean for girls was higher than the district mean for boys, for all grade levels. The district mean for boys was higher than the national mean for boys at all grade levels. The district mean for girls were very similar to the national mean for girls at all grade levels. The means, standard deviation and percent scores for the eleven content areas for both the district and national norm group are presented in Appendix A. The strengths and weaknesses of students were fairly consistent across both schools and grade levels.

The mean number of students in junior high groups was 95.3. The mean number of students in elementary groups was 33.6. Ten of the elementary estimates were based on a relatively small N (20 or less). The estimates based on a small N were consistent with the estimates based on more adequate N's.

The information collected was reported to principals, the superintendent's cabinet, central office consultants, health education teachers and the Board of Education.

Different reporting formats were developed for each group. The Board of Education, for example, received (1) a one page narrative, (2) a table of means and standard deviations for both the district and national norm group by grade by sex, and (3) a table of percent scores for district and national norm groups by grade by sex for each of the eleven content areas of the AAHPER Cooperative Health Education Test. A copy of the Board report is attached as Appendix A.

Principals, on the other hand, received detailed information on their schools. The school results were compared with local district norms and national norms. A copy of the report form used with building principals is attached as Appendix B.

The report to the superintendent's cabinet included:

1. School by school estimates of mean scores and standard deviation by grade level and sex.

2. School by school estimates of the percent of students scoring at or above the 75th, 50th and 25th percentile on both local and national norms by grade level and sex.

Health teachers, like principals, received primarily school level information, however, more emphasis was placed on item analysis information and on analysis of performance by the eleven content areas.

The reports represented the first time quantitative information was available on health education. The reports were well received by all groups. The information served to stimulate communication between groups interested in health education and was used to identify and

select inservice activities for health teachers throughout the year. The results were also used to help plan the content of a one week summer workshop for health education teachers.

As a result of the overall favorable reaction to the Spring 1975 Health Assessment, the assessment procedures were replicated in Spring 1976. Similar assessment procedures have been incorporated into the 1976-77 district testing program. The district health consultant now strongly advocates the use of test data in curriculum planning and is in the process of developing a physical fitness assessment.

Although the results of the 1976 Health Assessment have not been fully analyzed yet, building level reports have been generated.

Problems Encountered

The three most serious problems encountered during the assessment were:

1. Negotiation of contractual arrangements: A licensing agreement had to be negotiated which allowed the district to print the matrix test. A set fee of .06¢ was charged for each matrix test printed. Actually the problems encountered negotiating with Educational Testing Service were very minor, however; the contractual problems have in the past, made the use of the matrix sampling procedures very inconvenient.
2. Data coding problems: The use of matrix sampling procedures made it necessary for students to record the matrix test number (1-6) on their answer sheets. Since National Norms on the Cooperative Health Test are provided separately for boys and girls, sex also had to be coded on the answer sheet.

If either sex or matrix test number were missing, the student responses could not be used.

3. Small number of students: Since national norms were reported by sex, separate analysis had to be conducted for girls and boys. The number of students in some of the small elementary schools was not sufficient to make separate estimates of achievement for boys and girls.

None of the problems, however, are serious enough to preclude the use of the assessment procedures. For example, 96% of the answer sheets collected during the 1976 assessment were filled out correctly. This represents a slight increase over the 1975 assessment when approximately 95% of the answer sheets were useable. The 1976 scores of sixth grade girls at two schools could not be estimated because not enough students handed in "useable" answer sheets.[2] Scores were estimated for 108 of the 110 possible groups so failure to obtain estimates on two groups represents less than a two percent loss of information. Ten of 64 sixth grade estimates were based on less than 20 students, however, the estimates based on small number of students presented no serious problem.

Related to the problems of small number of students was the problem of negative variance. Negative estimates of the variance make it impossible to estimate the distribution. This problem was encountered twice in 1975 and twice in 1976 (or less than 2% of the runs).

---

[2]An arbitrary cut-off of 15 students was established for the 1976 assessment. One of the student groups that participated in the 1975 assessment, however, had only 9 students. Although evidence exists that supports the use of matrix sampling procedure with relatively small N's (20 students or more), the use of the procedure with very small groups is questionable.

Even the contractual problems encountered the first year were not serious enough to deter potential users. The 1976 assessment reused the matrix tests that were printed in 1975. Perhaps it would have been better to resample the items and build new matrix tests. This step would also necessitate printing new matrix tests and would increase the cost of the assessment. In the future, consideration will have to be given to reprinting the matrix tests each year. The eventual decision will have to be tied to both practical and logical considerations.

It is wasteful to limit the study to a 60 item pool. Multiple matrix sampling is an efficient technique for use with large numbers of items. Increasing the number of items would also increase the content validity of the assessment effort and provide more useable content information.

Educational Importance/Implication

The study demonstrated that multiple matrix sampling procedures can be used to conveniently collect district assessment data that otherwise could not or would not be collected. Building staff, who would have been reluctant to use an entire class period for testing, were willing to administer the tests at the beginning of an otherwise normal class period.

The health assessment data was subsequently presented to five different groups. In every case the information contained in the report was well received. All group information available from the traditional standardized testing program was provided through matrix sampling. In addition, information not normally provided in the testing program (school means, standard deviation, local normative information) was also provided.

The information provided by the assessment was used by the district health education consultant to plan curriculum changes and develop in-services sessions for health education teachers. The information served as a catalyst in discussing the health education curriculum and was the focal point of health education reports to the superintendent's cabinet, principals' council, consultants' council and the Board of Education.

The health education assessment has been incorporated into the normal district testing program, and matrix sampling procedures are being considered as alternatives to census testing for other district assessments.

APPENDIX A

LINCOLN PUBLIC SCHOOLS
INSTRUCTIONAL SERVICES

Report to the Board of Education December 16, 1975

HEALTH TEST RESULTS

The AAHPER* Cooperative Health Test was administered last spring in all Lincoln elementary and junior high schools. The test, which is published by Educational Testing Service, was first administered in the Lincoln School District by staff members of the Nebraska Center for Health Education, University of Nebraska-Lincoln, as part of a state-wide survey of health education at the 8th grade level. It was given in grades 6, 7 and 9 in Lincoln to collect additional information that would be useful in planning future health programs at both elementary and junior high school levels.

In order to save both teacher and student time, a sampling procedure was used in which each student answered only a few items. The whole process took about ten minutes of classroom time.

Results are summarized on the attached sheets. Scores are reported for the total test and on each of the eleven content areas. Key findings include:

1. As in national results, girls scored higher than boys at all grade levels. However, Lincoln boys scored higher than national norms for boys while Lincoln girls scored about the same as national norms for girls.

2. Lincoln boys scored higher than Lincoln girls on three of the eleven areas: community health, personal health care, and safety & first aid. Nationally, girls scored higher than boys on all eleven areas.

3. Lincoln boys scored highest in the areas of consumer health, nutrition, and community health, while Lincoln girls scored highest in the areas of nutrition, consumer health and growth & development

4. Lincoln boys scored lowest in the areas of personal health care, international health, and disease & disorder, while the Lincoln girls scored lowest in the areas of personal health care, international health care and mental health.

5. Areas judged to be of particular importance for which Lincoln students' scores were judged to be less than satisfactory were mental health, personal health care, and disease & disorder.

Dean Austin, Health Education Consultant
Carl Novak, Evaluator
Ron Brandt, Associate Superintendent for Instruction

---

*American Association for Health, Physical Education, and Recreation

Results of the Spring 1975 Administration of the AAHPER Cooperative Health Education Test

Mean Raw Score and Standard Deviation on the 60 Item Test

	Lincoln Mean	National Mean	School St. Dev.	National St.Dev.
GRADE 6				
Boys	33.3	NA*	10.3	NA*
Girls	35.3	NA*	9.2	NA*
GRADE 7				
Boys	36.8	33.7	9.8	12.1
Girls	38.7	38.7	8.8	11.1
GRADE 9				
Boys	43.8	42.0	9.3	11.1
Girls	44.9	44.8	8.1	9.0

*National Norms and Standard Deviations were not available for the 6th Grade.

Mastery Indices in Percent for Each of the Eleven Content Areas in the AAHPER Cooperative Health Education Test for the National Sample and the School District of Lincoln, NE by Grade by Sex

	GRADE 6				GRADE 7				GRADE 9			
	BOYS		GIRLS		BOYS		GIRLS		BOYS		GIRLS	
	Lincoln Mean	National Mean*	Lincoln Mean	National Mean*	Lincoln Mean	National Mean	Lincoln Mean	National Mean	Lincoln Mean	National Mean	Lincoln Mean	National Mean
Consumer Health	66	NA	70	NA	72	64	76	72	82	81	84	84
Community Health	57	NA	56	NA	68	66	65	68	80	80	79	80
International Health	44	NA	47	NA	51	43	60	56	66	63	67	63
Disease & Disorder	44	NA	50	NA	53	52	56	64	67	67	71	73
Personal Health Care	46	NA	47	NA	51	50	49	57	62	61	60	66
Sex Education	54	NA	56	NA	57	44	63	57	72	59	77	68
Growth & Development	58	NA	66	NA	67	57	74	70	78	69	82	79
Nutrition	64	NA	70	NA	71	67	78	76	79	78	84	85
Mental Health	53	NA	58	NA	59	55	59	61	64	65	64	69
Drug Use & Abuse	54	NA	56	NA	59	59	61	66	74	75	76	79
Safety & First Aid	63	NA	61	NA	61	50	65	55	72	67	70	64

*National norms were not available for 6th grade students.

APPENDIX B

RESULTS OF THE SPRING 1976 ADMINISTRATION OF THE AAHPER COOPERATIVE HEALTH EDUCATION TEST AT _____ JUNIOR HIGH SCHOOL

Mean Raw Score and Standard Deviation on the 60 Item Test

	School Mean	Lincoln Mean	National Mean	School St. Dev.	Lincoln St. Dev.	National St. Dev.
GRADE 7						
Boys	____	38.2	33.7	____	9.7	12.1
Girls	____	38.8	38.7	____	8.4	11.1
GRADE 9						
Boys	____	43.2	41.0	____	9.7	11.1
Girls	____	44.7	44.8	____	7.9	9.0

Percent of Local School Students Scoring at or above the 75th, 50th, and 25th Percentile Respectively when Compared to National Norms and Local Norms by Grade Level by Sex.

Percent of Students at or above the	School Results Compared With National Norms Boys	School Results Compared With National Norms Girls	School Results Compared With Lincoln Norms Boys	School Results Compared With Lincoln Norms Girls	Lincoln Results Compared With National Norms Boys	Lincoln Results Compared With National Norms Girls
GRADE 7						
75th percentile	____	____	____	____	28.4	22.8
50th percentile	____	____	____	____	62.2	53.5
25th percentile	____	____	____	____	89.9	87.9
GRADE 9						
75th Percentile	____	____	____	____	29.9	26.0
50th Percentile	____	____	____	____	53.7	50.6
25th Percentile	____	____	____	____	73.5	72.0

Mean Percent Scores for Each of the Eleven Content Areas in the AAHPER Cooperative Health Education Test for the National Sample, the School District of Lincoln, NE and the Local Building by Grade by Sex for _____.

GRADE 7

	BOYS				GIRLS			
	Number of Items	School Mean	Lincoln Mean	National Mean	Number of Items	School Mean	Lincoln Mean	National Mean
Consumer Health	5	_____	74	64	5	_____	77	72
Community Health	5	_____	65	66	5	_____	62	68
International Health	3	_____	53	43	3	_____	53	58
Disease and Disorder	5	_____	55	52	5	_____	57	64
Personal Health Care	7	_____	53	50	7	_____	51	57
Sex Education	6	_____	58	44	6	_____	66	46
Growth & Development	6	_____	67	57	6	_____	75	70
Nutrition	7	_____	73	67	7	_____	76	76
Mental Health	4	_____	60	55	4	_____	63	61
Drug Use & Abuse	8	_____	68	59	8	_____	62	66
Safety & First Aid	4	_____	64	59	4	_____	62	55

GRADE 9

	BOYS				GIRLS			
	Number of Items	School Mean	Lincoln Mean	National Mean	Number of Items	School Mean	Lincoln Mean	National Mean
Consumer Health	5	_____	80	81	5	_____	83	84
Community Health	5	_____	79	80	5	_____	74	80
International Health	3	_____	65	63	3	_____	66	63
Disease and Disorder	5	_____	66	67	5	_____	72	73
Personal Health Care	7	_____	61	61	7	_____	62	68
Sex Education	6	_____	70	59	6	_____	76	68
Growth & Development	6	_____	76	69	6	_____	84	79
Nutrition	7	_____	79	78	7	_____	84	85
Mental Health	4	_____	64	65	4	_____	68	56
Drug Use & Abuse	8	_____	71	75	8	_____	75	79
Safety & First Aid	4	_____	75	67	4	_____	70	64

MEASUREMENT PROBLEMS AND ISSUES RELATED TO APPLIED PERFORMANCE TESTING

JAMES R. SANDERS
WESTERN MICHIGAN UNIVERSITY

Dr. James R. Sanders
Associate Professor
Western Michigan University
Evaluation Center
College of Education
Kalamazoo, Michigan 49008

MEASUREMENT PROBLEMS AND ISSUES RELATED TO APPLIED PERFORMANCE TESTING

JAMES R. SANDERS

WESTERN MICHIGAN UNIVERSITY

(Summary)

Applied Performance Tests have been defined in Sachse and Sanders (1975) as "instruments designed to measure performance in an actual or simulated setting." They are measurement devices that require an actual or a close approximation of the setting to which the performance is expected to be transferred.

It is the thesis of this paper that the technical criteria used in the development and evaluation of applied performance tests are no different than those used for any other behavioral measurement devices. The unique aspects of applied performance measurement lie in the stress given to the degree of realism of stimulus and response conditions. Because criteria that are easily met with most psychological measures are not met with applied performance tests, some unique measurement problems arise. The purpose of this paper, after an initial excursion into the history and theory of applied performance testing, is to define relevant criteria used in evaluating applied performance tests and to discuss problems of applied performance testing associated with these criteria.

Two criteria used in the development and evaluation of any testing device are its reliability and validity; they are related. The reliability of a test places an upper limit on the criterion validity of that test. This holds true for any measurement device. Another way of looking at the relationship, however, is in terms of the setting in which the measure

---

Sachse, T. P. and Sanders, J. R. A Look at Applied Performance Testing in Education. Monograph of the Clearinghouse for Applied Performance Testing. Portland, Oregon: Northwest Regional Educational Lab, 1975.

is taken. Under tightly controlled conditions, the reliability of a measure can be very high, but at the price of its validity, assuming that the tightly controlled conditions do not reflect the ultimate criterion setting. As the testing situation becomes more real-life, the validity of the measure can increase, but usually at the cost of the control or reliability of the measurement. This trade-off is especially a problem when the real performance situation is the ideal because the reliability of measures taken under such "noisy" conditions is often quite low. Steps to deal with this problem are suggested.

Standardization of testing conditions is another testing criterion that is a problem area in applied performance testing. In real-life situations, standardized testing can rarely be achieved. Examples of standardized applied performance testing strategies are included in the paper as models for those who would use such testing devices.

Sampling errors produce another measurement problem with applied performance testing when the examiner wishes to generalize results to other settings, times, or persons. Idiosyncratic factors that can affect a person's performance in a real-life situation can come from many sources. A list of such factors is included in the paper.

Scoring problems associated with applied performance testing include all of those that are typically associated with observation as a means of data collection. Training observers to be sensitive to what to observe and how to record observations is critical. External distractions and the "halo effect" are two scoring problems that must be dealt with

by those who use applied performance tests. Strategies for dealing with these problems are discussed.

A set of problems associated with the cost of applied performance testing must also be mentioned when this testing approach is being discussed. One of the most serious criticisms of applied performance testing is that it costs so much in personnel time, facilities, obtrusiveness in normal operations, risk, and logistics--all of which serve to make its usefulness questionable. Costs can be reduced, as substitutions for the real-life situation are found, but measurement trade-offs, again, become a problem (e.g., reductions in criterion validity). Ways of dealing with the cost problem for assessmert purposes are suggested in the paper.

Because applied performance tests are often developed and used for unique purposes and settings, technical data on such tests are often not available. The test user typically does not have an available manual that contains information about the adequacy of the measurement device. This problem compounds all other measurement problems, and suggests that users should attend to, and plan to deal with the measurement problems outlined in this paper.

SYMPOSIUM ON:
LARGE-SCALE ASSESSMENT REPORTING AND USAGE:
DELAWARE AND GEORGIA AS EXEMPLARS

ROBERT BIGELOW AND HERVEY SCUDDER
DELAWARE DEPARTMENT OF PUBLIC INSTRUCTION
AND
EDUCATIONAL TESTING SERVICE

Mr. Robert Bigelow
Supervisor, Education Planning
State Department of Public Instruction
Townsend Building
Dover, Delaware 19901

Mr. Hervey C. Scudder
Program Director
Educational Testing Service
Princeton, New Jersey 08540

131

THE DELAWARE EDUCATIONAL ASSESSMENT PROGRAM
FROM THE PEOPLE -- FOR THE PEOPLE

ROBERT A. BIGELOW

DELAWARE DEPARTMENT OF PUBLIC INSTRUCTION

Introduction

Over the past five years, the Delaware Educational Assesament Program (DEAP) has become a recognized part of our state's educational scene. DEAP haa survived and grown in the midst of initial paranoia and frequent controveray. Unlike many other state assessment efforts, DEAP ia not legialatively mandated. Even so, Delaware citizens and educators have become very much aware of DEAP results and their implications. Each year when the assessment results are released, debates about the purpoaes and value of large-scale assesament are rerun in faculty lounges, local board meetings, college classes, and newspapers. And each year our State Department of Public Instruction beseeches the state legislature and searches its federal pocketbooks for sufficient funds to support "another year" of statewide assessment.

Of course, these problems are not unique to our state. But how were we able to survive our growing pains and atill provide relatively consistent assessment services to Delawareans? In the next few minutea, I would like to share with you two simple principles that do work in running our state assessment program: (1) assessment for improvement; and (2) acceptance through involvement.

Assessment for Improvement

In the late 1960's, interest in legislating accountability through mandated testing programs was sweeping the country. At this time, most Delaware educators were expressing their concern that accountability through test scores alone would have serious implications (which many of us have subsequently witnessed). For these reasons, the DEAP was implemented as part of a long-range plan for the improvement of state and local educational programs. Activities implemented through this plan are directed toward answering four major questions:

* What do we want from our educational system?
* What have our students attained?
* What are our program strengths and weaknesses?
* What can be done to improve our educational programs?

As indicated by the above questions, the formulation of statewide goals and objectives was prerequisite to the implementation of needs assessment activities. By 1971, statewide learner goals for the 70's and 80's were established. So far, statewide educational objectives related to these broad goal areas have been developed for the content areas of reading, English/language arts, mathematics, science, social studies, and mental and physical health. The objectives have been disseminated to educators throughout the state in order to facilitate curriculum efforts for kindergarten through grade eight students. Terminal objectives for secondary students are currently being developed based upon recent interests in survival skills and minimal requirements for high school graduation.

In 1971, the DEAP was initiated by the Planning, Research, and Evaluation Division of the department to provide state and local educators with information to answer the second and third major questions pertaining to

student attainments and program needs. Over the past five years, norm referenced survey batteries have been administered to all regular first, fourth, and eighth grade students throughout the state. In addition to student ability and achievement data, information about community and school resource factors has also been collected.

Each year DEAP results have been generated and distributed in various ways to students, teachers, educational administrators, legislators, and the community at large. Each year local officials have maintained the responsibility for disseminating their district and school assessment results in a manner they felt appropriate. However, from the very beginning, the expected misuses of the data occurred: unfair achievement comparisons were made between schools and districts, high scoring districts were cited as "the place to live", and teachers began to be concerned about the effect of test scores upon contract negotiations. In efforts to dissipate these concerns, the department has focused all assessment reports, field services, and publicity on the primary purpose for the DEAP -- to generate information about local and statewide educational needs as a basis for future program improvement.

Over the years, obvious misuses of the assessment results seem to be diminishing. To the best of my knowledge, there has been no mass movement to high scoring districts and no teacher has been fired because of assessment results. On the contrary, most teachers participating in the DEAP have had multiple opportunities to understand DEAP results for their school and district. Local school board members and chief school officers annually review their district achievement results in light of community and educational resources. Newspaper articles have begun to report district programs being implemented to alleviate curriculum weaknesses indicated by the survey tests. Finally, an increasing number of project applications utilizing DEAP results in their needs assessment sections are being generated each year. These

proposals represent millions of dollars that are being allocated yearly to local districts to support programs meeting educational needs indicated by the DEAP results.

Acceptance Through Involvement

Another reason that DEAP enjoys increasing acceptance is that each phase of the program is largely determined through concrete inputs from those it serves. Examples of the important kinds of inputs that DEAP participants contribute to each of the major assessment activities are discussed below.

Refinement of learner objectives. Every year representatives from local school districts, higher education, and the department have been asked to participate in the continued refinement of statewide educational objectives. These people are organized into subject area task forces whose mission is to annually revise and/or rewrite the existing statewide educational objectives. This process can take many weeks since preliminary drafts of these objectives must first be reviewed by local educators before final consensus is reached through a modified Delphi technique. Even though the statewide objectives are still incomplete, the yearly outcomes are of primary importance -- updated sets of objectives developed and approved by Delaware educators.

Development and/or revision of assessment instruments. The DEAP batteries used to collect information about student ability and achievement are also annually reviewed. Instrument development and/or revision is based upon inputs from three sources: (1) item specifications recommended by the subject area task forces; (2) teacher comments about test content collected during test administrations and inservice sessions; and (3) professional contributions from our contractor, Educational Testing Service. In one case, Delaware fourth and eighth grade social studies tests were constructed, piloted, and refined through the joint efforts of social studies task force members, reading specialists, department staff, and ETS. This year, selected members from each

subject area task force will be trained to generate their own item alternatives to specifications recommended during a summer workshop. ETS staff will then review these items and recommend final selections for inclusion in the revised tests. In this way, DEAP participants should feel even stronger ownership of the revised assessment instruments.

Administrative logistics. Managing the flow of assessment materials becomes more efficient with each year of the program's operation. There is not enough time here to adequately describe the kinds of cooperation among our district test coordinators, which has enabled us to deliver, administer, assemble, ship, and score assessment materials for every school district within 20 working days. ETS has also provided valuable assistance in this effort as will be described during the next presentation. Rising operating costs have recently forced the department to take over many of the logistical activities previously conducted by ETS. However, we have still been able to meet our timelines because of successful management patterns developed jointly with our contractor and the continued support of our local test coordinators.

Reporting assessment results. DEAP results are distributed to and utilized by many audiences. Our primary targets have always been classroom teachers, curriculum managers, and principals of each participating school. Through mini projects implemented in each school district, these people learn to interpret the assessment results in light of their own curriculum outcomes and actually use the data to identify further areas of investigation and improvement. In turn, demands for more specific information about curriculum outcomes and corrective action have generated local requests for in depth assistance from the department's dissemination unit and field service staff. At another level, results from a longitudinal investigation of school resources and achievement has prompted a request by the State Board of Education for an in depth follow-up study of schools performing above or below expectations.

Finally, analyses of schievement results have indicated increasing differences between Delaware and national norms as students progress through the elementary and middle grades. This downward trend has been of concern to many educators and citizens throughout the state. These concerns are being voiced in professional and public demands to expand DEAP services to provide more information about student attainment of each statewide objective.

The Future of DEAP

We are initiating a new phase of our assessment program. Every two or three years the traditional norm referenced batteries will be cycled in at each participating grade level for benchmark studies of long-term changes. In addition, we plan to annually administer objective referenced measures in selected content areas beginning next fall. We anticipate that these new measures will supply more complete information about student needs in each subject area assessed. This additional information should also reinforce further diagnostic efforts at the local level to help practitioners better address the final and most important question of our long-range plan (What can be done to improve educational programs?).

We are later than other states in entering the arena of objective referenced measurement. However, we are finally ready; our people understand it, and our people want it.

SELECTED BIBLIOGRAPHY

Additional papers, handouts or brochures which were not submitted or presented as part of the formal paper sessions were made available to participants during the three-day program. Because these materials cover topics of general interest to assessment personnel they are being referenced here so that interested readers will be aware of them.

Free single copies of the materials titled below can be obtained by writing directly to the author/ presenter.

LINCOLN PUBLIC SCHOOLS' POSITION PAPER ON ASSESSMENT

Dr. Ron Brandt
Associate Superintendent for Instruction
Lincoln Public Schools
PO Box 82889
Lincoln, Nebraska 68501

THE USE OF CORRELATES OF ACHIEVEMENT:
CHANGEABLE AND UNCHANGEABLE

Dr. Paul Campbell
Director, ESS
Educational Testing Service
Princeton, New Jersey 08540

SUCCESSFUL UTILIZATION OF STATE ASSESSMENT RESULTS BY LEAs

Dr. J. Robert Coldiron
Chief, Division EQA
State Department of Education
Box 911
Harrisburg, Pennsylvania 17126

TEXAS CAREER MEASUREMENT SERIES:
FROM ASSESSMENT TO INSTRUCTION

Mr. Keith Cruse
Program Director, Assessment
Texas Education Agency
201 East 11th Street
Austin, Texas 78701

1975-76 MICHIGAN ASSESSMENT PROGRAM: PARENT REPORT
AND CITIZENS GUIDE PROJECT REPORT

Mrs. Judith E. Moyer
Education Research Consultant
State Department of Education
PO Box 420
Lansing, Michigan 48902

DISTRICT USE OF ASSESSMENT RESULTS

Dr. Alan Robertson, Chief
Division of Research
Bureau of Occupational Educational Research
State Education Department
Albany, New York 12234

A METHOD FOR EVALUATING ASSESSMENT

REPORTING THE RESULTS OF STATEWIDE ASSESSMENT

SETTING STANDARDS AND LIVING WITH THEM

Dr. Lorrie Shepard
Laboratory of Educational Research
University of Colorado
Boulder, Colorado 80302

OAKLAND MICHIGAN TESTING HANDBOOK

Dr. Richard Watson
Oakland Intermediate School District
2100 Pontiac Lake Road
Pontiac, Michigan 48084