DOCUMENT RESUME

ED 104 948 TM 004 406

AUTHOR Hall, Mary
TITLE Statewide Assessment of Student Performance: A Comparative Survey.
PUB DATE [Apr 75]
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D. C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS Academic Achievement; Achievement Tests; Comparative Analysis; Decision Making; *Educational Assessment; Educational Objectives; *Educational Status Comparison; Elementary Secondary Education; Evaluation Methods; Measurement Techniques; Methods Research; *National Surveys; Program Effectiveness; Program Evaluation; Standardized Tests; *State Departments of Education; *State Programs; Student Evaluation

ABSTRACT

A survey of 42 statewide assessment programs was conducted to determine: (1) The status of statewide assessment programs in the United States when classified by purpose, authority, methodology, and scope; (2) Are there any differences within these classifications for programs which are aimed primarily at state-level decision making as opposed to those designed primarily for local use; and (3) The primary types of measurement used by statewide assessment programs and the strengths and weaknesses of such models. Data was collected by requesting 53 state departments of education to send information and publications related to their statewide assessment activities. Materials received were checked against two nationwide descriptions of state assessment and or testing programs issued by Educational Testing Service in 1973. Some recommendations for future research include: the need for immediate research on the question of the most effective roles for statewide assessment programs in influencing state or local decision making, research needed on the procedures and techniques to widen availability of criterion referenced instruments, and research studies that will solve some of the methodological problems facing state assessment programs. (Author/DEP)

STATEWIDE ASSESSMENT OF STUDENT PERFORMANCE: A COMPARATIVE SURVEY

Mary Hall

Assistant Superintendent

Oregon State Department of Education

Salem, Oregon

Presented at

American Educational Research Association
National Meeting
March 30 - April 3, 1975

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION

A. Need for the Study

In the late 1960's, the nation's educational community began to hear a new term, "accountability." Since that time, this phrase has been loosely applied to a variety of philosophies and methodologies. But as coined by Leon Lessinger, the term was associated with a method trying to insure what he called the "three basic rights in education." In discussing Dr. Lessinger's approach, John Brademas noted:

> "The first right is the child's; the second, the taxpayer's; and the third, the school's. The child's right, said Dr. Lessinger, was 'to be taught what he needs to know' in order to be a productive and satisfied member of our society; the taxpayer's right, to be informed of the educational results produced by specific expenditures; and the right of the school, finally, was to draw on all the resources of the society..." (1)

In meeting the goal of guaranteeing all children the right to acquire certain basic skills, Dr. Lessinger proposed a methodology which he called "an independent educational audit of educational results." While a variety of specific approaches has been suggested or adopted for implementing this methodology, one of the most consistent and pervasive has been the initiation of statewide programs of student assessment, that is, programs designed to measure, on a statewide basis, pupil achievement in designated priority areas.

The spread of such programs has been phenomenally rapid and complete. According to Educational Testing Service's Center for Statewide Educational Assessment, all 50 states plus the District of Columbia, the Virgin Islands and Puerto Rico have reported assessment activities either as operational, in a developmental process or in a planning stage. (2)

The speed with which states have responded to the methodology of "accountability" has brought corresponding problems. As House, Rivers and Stufflebeam point out,

> "Educational accountability is not easy to work with. Educational researchers have not produced tested standards and procedures for state accountability systems. But many states' legislatures have mandated that systems of accountability be implemented on a crash basis." (3)

Gaps in the research base supporting statewide assessment programs are a critical problem. "Warehousing" of state assessment models and procedures by the ETS Center for Statewide Educational Assessment and the State Education Accountability Repository operated by the Wisconsin State Department of Education has led to a number of descriptive publications in the field.

But as will be noted elsewhere in the paper, little research and development at the national level appears to be underway in major categories such as the development of alternative modes for assessment instrumentation, methodology for achieving statewide consensus on the purposes and focus of assessment, the role of assessment data in state or local policy development, more effective techniques for dissemination and utilization of assessment results and other areas of study essential to improvements in the design and operation of statewide assessment programs.

It is hoped that this paper will be one step towards filling these gaps by providing a comparative survey of the status of selected aspects of statewide assessment programs as of the summer of 1974 and by suggesting several dimensions for needed research activity in the future. In addition, by highlighting some of the major differences between such programs, this study may be of use to state assessment directors in challenging their assumptions as to why and how they are operating the type of program chosen for implementation. Finally, the data may be of use to those few states who have not yet proceeded to the final implementation stage and who may yet

have an opportunity to compare their planning against national patterns in key areas such as the purposes for assessment programs, the types of tests used, the domains selected for testing and so forth.

B. Statement of the Problem

The purpose of this paper is to answer the following questions: What is the status of statewide assessment programs in the United States when classified by purpose, authority, methodology and scope? Are there differences within these classifications for programs which are aimed primarily at state-level decision-making as opposed to those designed primarily for local use? What are the primary types of measurement used by statewide assessment programs and what are the strengths and weaknesses of such approaches?

C. Assumptions and Definitions

As Frank Womer (4) found, there are a variety of definitions for statewide assessment currently in use. For the purpose of this paper, statewide assessment of student performance has been defined as "a comprehensive effort to gather statewide information on the status of student progress towards desired goals and/or objectives for the purpose of educational decision-making."(*) Given this definition, such decision-making may take place at either the state or local level. The significant variables are that the program be comprehensive, that it focus on statewide information, and that this information include student performance data on desired goals and/or objectives.

Since much of this study is based on information found in documents published by State Departments of Education which describe their statewide assessment program, it is assumed that such publications represent accurate descriptions of what is taking

---

(*) This definition is a modified version of the description adopted by the Oregon State Department of Education for its assessment programs. A report on the two year study which led the agency to its definition is contained in Indicators and Statewide Assessment, Cooperative Accountability Project, Denver, Colorado, March 1974.

taking place in each state. While some verfication was obtained when other sources describing specific state programs appeared to conflict with the states' reports, there was no attempt made to observe assessment operations personally in each state or to contact other agencies for an independent report on program details.

D. Limitations and Related Research

Given the national scope of this study, several limitations were adopted in order to keep the paper within manageable bounds. If literature was available on certain aspects of assessment, it was assumed that these topics did not represent gaps in the awareness of researchers about the need for study in this area. Generally, the availability of such writings was used as a major criterion for delimiting the scope of the study. For this reason, the limitations of the study are presented in concert with a review of related research.

One cautionary note should be observed. The availability of writings on a particular topic does not necessarily mean that research per se has been conducted in that area. Although the author gave only a cursory review to articles outside the primary focus of the paper, a general impression was gained that most of the writings appear to be based on philosohical or emotional concerns as opposed to empirical results. The whole dimension of accountability and assessment still remains substantially open to the educational research community.

The first limitation adopted was the decision not to argue either the disadvantages or advantages of the accountability movement. Readers interested in this subject will find an overwhelming amount of literature available through ERIC or other bibliographic sources.

Second, the study does not intend to assess the merits of the National Assessment of Educational Progress nor debate the pros and cons of state assessment programs as a

general phenomena. The former area is covered by such writers as Beymer (5), Findley (6), and Conaway (7), with the latter field described by authors such as Buchmiller.(8) Again, a cursory review of literature along this dimension revealed many articles, but only a few appear to be directed towards experimental or evaluative results.

Third, the paper does not provide an indepth review of the legislation establishing state assessment programs. An annual survey of this nature has been available since 1972 under the auspices of the Cooperative Accountability Project.(9)

Fourth, the study is not intended to provide evaluation of any single state assessment program. In fact, all findings in the study are presented as summary results and variables related to specific states are not identified as such. Evaluation of statewide assessment programs is a recently emerging field in educational research and has been a topic proposed for study at meetings of the American Educational Research Association. For those interested in this field, one of the first such studies given national distribution is the review of the Michigan assessment program by House, Rivers and Stufflebeam.(10)

Fifth, the study does not examine the types of information collected by state assessment programs on student, school or community variables for the purpose of helping to explain student testing results. According to a recent survey by Educational Testing Services' Center for Statewide Educational Assessment, (11) all states collect and report such correlates. The data is so extensive that it would warrant a separate study in its own right. Preliminary reviews of the extent of variables collected and their potential use for decision-making have been provided by Campbell (12) and the New York State Department of Education. (13)

In addition to the limitations suggested by the availability of related literature, two other decisions were made which affect the dimensions of this study.

First, states were not included in the reported results unless they had an assessment program actually in operation by spring of 1974. This excluded some nine states still in the planning stage.

Second, states were eliminated if a review of their program showed that it did not fit the definition of "assessment" as presented earlier in this paper. That is, the program must be comprehensive, focus on statewide information and include student performance data related to desired goals and objectives. Two states were found to meet the first two criteria, but to not collect data on student performance. Instead, they carried out their assessment by asking various categories of people what they thought should be the focus for state educational efforts. No effort was made by these states to initiate measurement activities to see if these expectations were actually being achieved. These two states were subsequently eliminated from the findings.

E. Summary

This chapter has attempted to document the purposes behind the study, to outline the statement of the problem, to provide definitions and assumptions and to describe the limitations of the study primarily through a review of related research. Chapter II shall describe the procedures used to collect data and the categories of program dimensions included in the analysis. Chapter III will review the findings of the paper, with Chapter IV focusing on recommendations for future research.

CHAPTER II: PROCEDURES

Data collection for this paper was initiated in the spring of 1972 with a personal letter to the planning and evaluation director of each State Department of Education throughout the country and to corresponding officials in the District of Columbia, Virgin Islands and Puerto Rico. These individuals were asked to describe their assessment program (or plans) and to furnish copies of any documents available on the state assessment effort. This initial survey of materials was incorporated into a preliminary study analyzing eight state assessment models. (14) Part of this work has been incorporated into the conceptual base of the current investigation.

This initial survey was followed by a similar letter to each state department in the winter of 1974. In all, some 300 documents were received and reviewed as a basis for the current analysis.

Materials received from each state were checked against the Educational Testing Services' 1973 survey of state assessment programs (15) and national review of state testing programs. (16) Where discrepancies occurred in the material furnished by the state for these two national publications and the documents furnished to the author, a personal phone call was placed to the state assessment director or planning and evaluation administrator. In all, thirteen states were contacted by phone during the late spring and early summer of 1974 to verify program details.

Four descriptive categories were developed for analysis of the materials: purpose and use of the program; basis of authority; type of measurement and population design; and scope of the program.

Within each category, several questions were posed dealing with aspects of assessment programs thought common to most states. Some questions were suggested by variables reported in the 1973 ETS Survey of Statewide Assessment Programs (17) and/or the 1973 Review of State Testing Programs. (18) Other questions (such as those dealing with the uses of assessment data and the types of measurement) were chosen by the author in the hopes that such information might provide clues for future research of the type desired by House, Rivers and Stufflebeam. (19)

All of the questions were stated as explicitly as possible in order to allow for unambiguous classification during the review of state materials. In most cases, the phrasing of the questions led to a "forced choice" response which limited the degree of subjectivity involved.* Table I identifies the analysis questions contained in each of the four categories.

Each state's material was then reviewed and a sheet was prepared for each recording how that program corresponded to the analysis questions. As noted earlier, 11 of the 53 states and territories originally contacted were excluded because they did not meet the study's criteria. The results from the remaining 42 states were then tabulated and summarized, with findings reported in Chapter III.

A second round of review was then undertaken to separate the responses to analysis questions of those states whose primary purpose was to produce data for state-level decision-making as opposed to those whose primary purpose was to generate information for use by local, participating schools. After compiling the data for each category, Fisher's Z Test of Significant Differences for Uncorrelated Proportions was applied to

---

* In most cases, the analysis questions could be answered with no ambiguity by referring to explicit statements in the state's own publications or to information provided by the state to the two national surveys cited earlier. In those few cases where the answer was not obvious, the author contacted the state's assessment personnel by phone and asked how they would classify their program. The validity of the data is thus assumed to be fairly high.

TABLE I: ANALYSIS QUESTIONS BY CATEGORY*

CATEGORY	ANALYSIS QUESTIONS
I: PURPOSE OF PROGRAM	Is the primary purpose of the program to furnish data for state-level decision-making or for decision-making by local, participating schools?
USES OF THE DATA PROVIDED BY THE PROGRAM	Does the assessment program propose to measure progress towards state educational goals adopted by the State governing board or legislature?
	Does the assessment program propose to predict an expected level of performance for students in participating school districts and then report actual results as a comparison to this prediction?
	**Is the data from the assessment program reportedly used for selecting priorities which are then used as the basis for allocating state or federal funds?
	Is the assessment program intended to produce information which is used to compare performance between districts?
	Is the data from the assessment program reportedly used to establish priorities for the allocation of funds under Title III of the Elementary and Secondary Education Act.
II: BASIS OF AUTHORITY	Is the state's assessment program mandated by the State Legislature?
	Does the state education agency (or state law) require local school districts to participate in the program?

* Two additional important categories of questions were originally considered for the study. These were questions dealing with the administrative design for such programs and questions dealing with the analysis and reporting designs. A cursory review of the state assessment documents showed such variation in whether this type of information was made available (and if so, in what detail) that the author decided not to attempt analysis along these two dimensions. Both topics warrant additional studies and include a variety of problem areas of concern to current assessment administrators.

** This question exempted resource decisions made under Title III of the Elementary and Secondary Education Act as this variable was specifically identified in another question.

TABLE I: Continued

CATEGORY	ANALYSIS QUESTIONS
III: TYPES OF MEASUREMENT	Does the state rely primarily on the administration of commercially-available, norm-referenced, standardized tests?
	Does the state rely primarily on the assessment exercises developed by the National Assessment of Educational Progress?
	Does the state rely primarily on criterion-referenced instruments other than those provided by NAEP?
	Does the state rely on an "eclectic" measurement model, using a mix of norm-referenced, objective or criterion-referenced tests, or other sources of data on student performance?
POPULATION DESIGN	Does the state test all children at specified grade levels?
	Does the state test a sample of children at specified grade levels?
	Does the state use a "mixed" design, testing all children in some grades and a sample of students in others?
IV: SCOPE OF PROGRAM	At what age or grade levels does the state administer tests?
	Does the state test only in the cognitive domain?
	Does the state test only in the affective domain?
	Does the state test in both the cognitive and affective domains?
	In what cognitive areas does the state administer tests?
	In what aspects of the affective domain does the state administer tests?
	Does the state test in the psychomotor domain?

determine whether any of the differences could be classified as "statistically significant" (assuming a ratio of 1.96 at the .05 level of confidence). These findings are explained in Chapter III.

Finally, the author divided the state score sheets into the four major types of measurement contained in category III and reviewed the state response for each. A major purpose of this review was to determine whether the type of measurement chosen appeared to be producing the type of data needed by the state in meeting its formal purposes for initiating a statewide assessment progra. From this analysis (and narrative information provided in the state documents), a statement on the strengths and weaknesses of each measurement type was generated. It should be recognized that this section of the paper represents the author's own viewpoint and the results may or may not be compatible with the opinions of statewide assessment personnel.

This chapter has outlined the methods used to collect and analyze data. Chapter III will report on study findings.

CHAPTER III: FINDINGS

This chapter is divided into three parts corresponding to the three questions listed in the statement of the problem. Part I discusses findings relative to the status of statewide assessment programs in the United States when classified by purpose, authority, methodology and scope. Part II discusses whether differences occurred in these classifications for states whose primary purpose was to produce information for state-level decision-making as opposed to those primarily serving the decision-making needs of local schools. Part III looks at the primary types of measurement used by state assessment programs and attempts to identify some of the strengths and weaknesses of each type. A summary of findings will be presented at the end of each part.

Part I: Status of Programs by Four Major Classifications

As noted in the discussion on study procedures, twenty-two analysis questions were generated and divided into four major classifications for the purpose of determining the current status of statewide assessment programs. Each of these will be discussed in turn.

Purpose and Use

Of the 42 states included in the study, 28 states (or 67%) of the population indicated that their primary purpose was to develop information to be used for decision-making at the state level. The remaining 14 states conducted their statewide programs primarily to produce data of use to local, participating schools.* Findings relevant to the uses to which states put their assessment information are shown in Table II. It is interesting to note that except for one analysis question, less

* The ETS Center survey (20) in 1973 showed exactly the same pattern even though that report included 24 states which were still in a planning mode. Apparently, those states which actually implemented their program during 1974 were consistent in sticking to the purposes indicated a year earlier.

TABLE II: USES OF STATEWIDE ASSESSMENT PROGRAMS

ANALYSIS QUESTION	PERCENT OF 42 STATES ANSWERING "YES"
Does the assessment program propose to measure progress towards state educational goals adopted by the State governing Board or Legislature?	29%
Does the assessment program propose to predict an expected level of performance for students in participating school districts and then report actual results as a comparison to this prediction?	10%
Is the data from the assessment program reportedly used for selecting priorities which are then used as the basis for allocating state or federal funds? *	14%
Is the assessment program intended to produce information which is used to compare the performance of the districts with each other?	17%
Is the data from the assessment program reportedly used to establish priorities under Title III of the Elementary and Secondary Education Act?	69%

*This question exempted Title III ESEA allocation decisions since that specific program was identified in a separate question.

than 30% of the states responded "yes" to any of the variables chosen for study. Why?

An obvious response might be that the wrong questions were asked, or that those included were stated in the wrong way. On the other hand, this set represents a fairly common list of the types of uses which are frequently cited as justification for initiating assessment programs. The general impression is given that while states may be clear on why they initiated the program (legislative pressure, federal requirements, response to the "accountability" movement), they are less clear about the specific types of decisions they will make once the information is available. House, Rivers and Stufflebeam (21) made this same observation in their evaluation of the Michigan state assessment program when they found difficulty in identifying decisions that resulted from the availability of the data. "Perhaps our most unexpected finding is that the assessment program has little apparent value for any major group," (22) they concluded.

Several other observations on the data can be made. A significant number of the 42 states (69%) reported that they are now using their statewide assessment program for the purpose of establishing priorities for funding as required under Title III of the Elementary and Secondary Education Act. This appears to be a positive step since it indicates that many states have chosen not to finance (and subject schools and students to) a separate type of assessment activity to meet this federal requirement. In addition, the discretionary nature of Title III ESEA is intended to allow states to use the money to meet needs specific to their locale, and problem areas identified through state assessment programs are a logical focus for the development of improved management or instructional techniques.

However, once Title III, ESEA is exempted as a type of resource-allocation decision, only 14% of the states provided evidence of making any other types of decisions on the distribution of state or federal funds as a result of the assessment data. The percentage of states using assessment data for resource allocation decisions might have been higher had the criteria for this question not been that states must either report making a specific type of allocation decision (such as setting priorities for Title I of ESEA or awarding state funds to districts whose assessment scores showed a certain pattern) or showed evidence of having adopted internal decision-making machinery using assessment data for the allocation of funds either to programs within the State Department or to local schools. This area remains one in which further study is needed.

Discrepancies on the specific focus of decision-making occurred in eight states who specified resource-allocation as an outcome of assessment on the material submitted for the ETS Center survey. (23) However, publications issued by the states themselves did not refer to this type of use. Phone calls to these states indicated that while they "intended" to use their assessment data in this manner, they had not yet built any internal procedures to assure that it happened. They were hoping that the assessment data would be useful, but they had no specific plans for insuring that the information impacted on appropriate decision-makers.

The relationship between assessment results and actual decision-making is a crucial factor in the eventual viability of such programs. The factors which have either supported or inhibited the effective use of assessment data need to be identified and models and procedures developed which other states may adopt to increase the effectiveness of their investment in assessment.

Basis of Authority

Two questions were posed in this category-one dealing with whether the program had been mandated by a state legislature and the other focusing on whether the state (either by law or by State Department regulation) required participation in the program by local school districts.

On the first dimension, it was found that 12 states (or 29%) of those actually having initiated a program by spring of 1974 had done so under statutory requirement of their legislature. Thirteen states (31%) required participation by local schools. It was not determined whether this participation was mandated by the legislature or was simply a state agency regulation. The sanctions against districts which chose not to participate were also not specified.

Methodology

Two categories of questions were generated along this dimension - one set dealing with the type of measurement adopted by the state and the second dealing with the population design selected. Based on the author's previous study, (24) four primary types of measurement were selected for analysis and are defined below along with the percentage of states relying primarily on the use of that approach.

Standardized	Student performance results are secured through administration of standardized, norm-referenced test instruments available from a commercial publisher;	52%
National Assessment	Student performance is determined through the use of assessment exercises developed by the National Assessment of Educational Progress (NAEP);	14.5%
Objective Reference	Student performance is determined through the use of objectives selected by the state and measured by instrumerts designed to show the number of students scoring correctly on the objectives. States relying primarily on the NAEP material (which is also a type of objective reference test) are excluded from this category;	28.5%

Eclectic	Student performance is not judged against a single source of objectives, but data is collected through a variety of means including student scores on tests administered within the state for other purposes. Both norm-referenced and objective reference instruments may be used.	5%

It is interesting to note that when both the National Assessment and Objective Reference types are combined, 43% of the forty-two states are now using some type of objective-reference instrument as their primary method of measurement. This is up slightly since 1973 when only 40.5% of the states reported the use of such instruments on the ETS Center survey. (25)

In looking at the question of population design, the study was expected to show whether states tend to use a sampling procedure in their assessment program, test all of the children in the state at specified grade levels, or use both a sample and total population approach.

The findings indicate that the majority of states (64%) rely on a sampling approach; 31% test all of the children in specified grade levels and 5% use a combination of the two designs.

Scope of the Program

Three types of questions were posed within this category dealing with the age or grade level of students tested; the domains (cognitive, affective, or psychomotor) within which states test; and the specific subject matter included in state assessment programs.

It was found that most states tend to design their programs by grade level (as opposed to student age) so all state findings were converted to a grade basis (in areas of doubt, a phone call was placed to the particular state to determine the

grades in which the majority of their students were tested). The findings for this dimension are reported in Table III. The most popular grades for testing were found to be grades 4, 8 and 11. In justifying these decisions, many states alluded to the fact that they did not have sufficient resources to test at all grade levels and thus preferred to design their assessment program to secure a status report on student performance at periods normally corresponding to the end of primary instruction, the beginning of high school and the termination (or close to termination) of secondary schooling.

In looking at the domains within which states tested, it was found that:

41 percent of the states measure only within the cognitive domain;

7 percent test only in the affective domain;

43 percent assess in both the cognitive and affective domain;

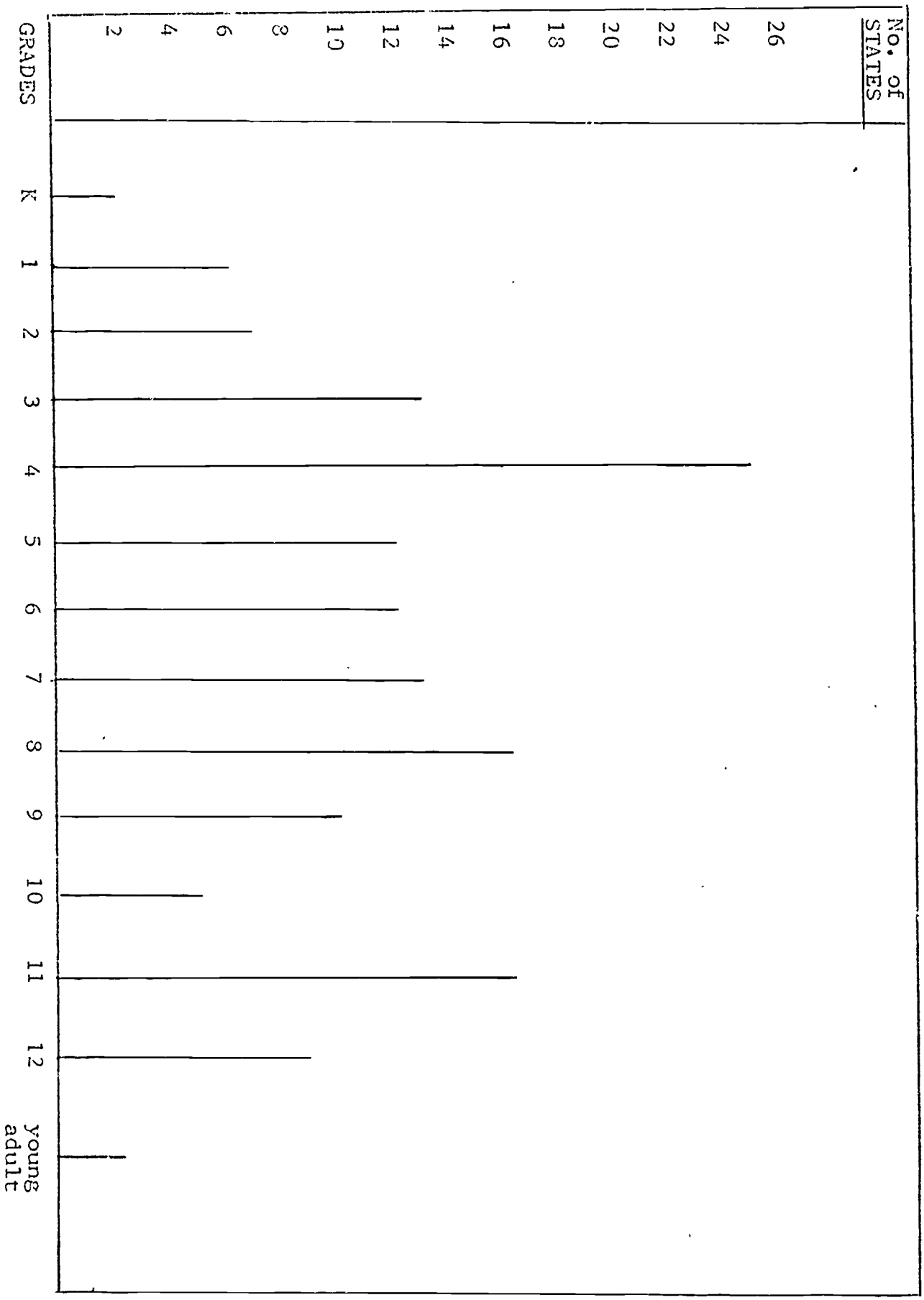
2 percent test only in the psychomotor domain;

7 percent assess student performance in all three domains.

Thus the major focus of state assessment programs is to test in both the cognitive and affective areas, with assessment only of cognitive skills running a very close second.

A third category of questions within this classification dealt with the specific subject matter in which tests are administered. In looking at the affective area, it was found that the majority of states focused on student attitudes towards school, with attitudes towards self, home or community running significantly behind. No breakdown was available on skill areas assessed in the psychomotor domain.

TABLE III: NUMBER OF STATES TESTING AT EACH GRADE LEVEL



In the cognitive area, it was found that states conduct testing programs in some 18 separate fields, with mathematics and reading by far the most popular. The results of this finding are presented in Table IV.

Summary of Part I

Part I of this chapter has reported the findings on twenty-two analysis questions designed to determine the current status of statewide assessment programs when classified by purpose and use, authority, methodology and scope. Significant findings include:

- The majority of states have initiated these programs to produce information for state-level decision-making;
- The majority of states use their assessment programs to establish priorities for the allocation of funds under Title III of the Elementary and Secondary Education Act.
- Less than a third of the states report making decisions as a result of their assessment data in other areas of use frequently cited as justification for the initiation of statewide assessment programs.
- Twelve states (29%) have assessment programs mandated by state legislatures and 31% of the states require participation of local schools.
- The majority of states (52%) rely on a standardized, norm-referenced type of measurement in their assessment program, although 43% of the states rely primarily on some type objective-referenced instrumentation.
- The majority of states (64%) use a sampling approach in selecting their student population, with 31% testing all children in the state in specified grades.
- States test pupils in all grades running from kindergarten through samples drawn to represent "young adults." However, the preference tends to be for assessing student performance in grades 4, 8 and 11.
- States test pupils in all three of the domains - cognitive, affective and psychomotor. States preferred to test either in both the cognitive and affective areas (43%) or cognitive only (41%).
- Assessment programs are operating in 18 subject matter fields within the cognitive domain, with heavy preference shown for reading (81% of states test in this field) and mathematics (78% of states assess in this subject).

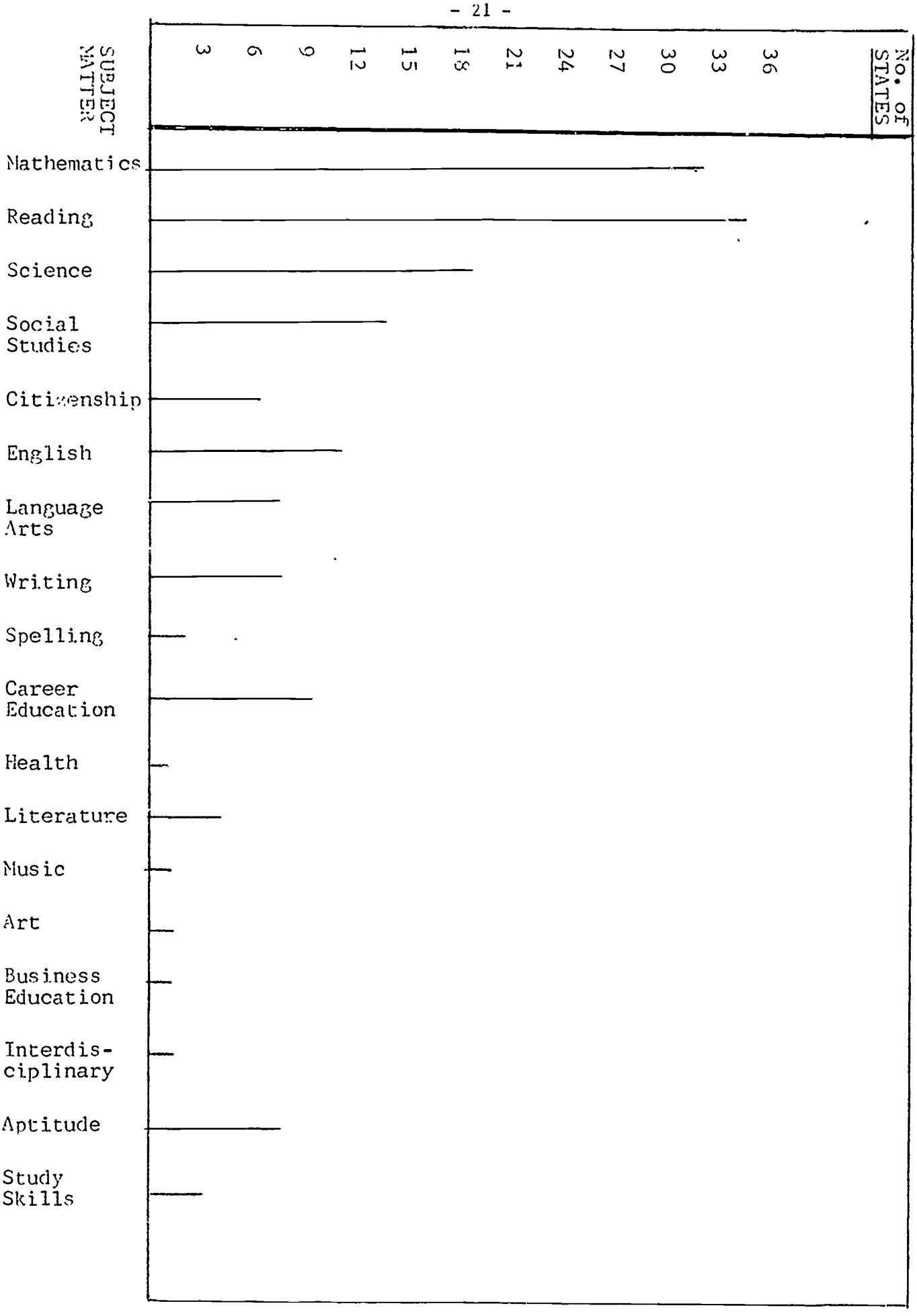


TABLE IV: NUMBER OF STATES TESTING IN EACH SUBJECT MATTER

Part II: Differences Between State-Oriented and Locally-Oriented Programs

As reported in Part I, 67% of the 42 states reported that the major focus of their program was to serve state-level informational needs, while the remaining states indicated their primary purpose was to produce data of use to local, participating schools.

Is this difference in purpose also reflected in results obtained by the two groups on other analysis questions? The answer is frankly dependent on whether the reader wants to insist that such differences be "statistically significant" for a study involving only 42 participants.

While "actual" differences were observed on almost all of the analysis questions, only one of the items (the question dealing with types of measurement) turned out to be statistically different at the .05 level of confidence when correlated using either Fisher's $\hat{z}$ Test of Differences for Uncorrelated Proportions or chi square analysis. Since techniques for correlating results for such small samples are limited, the "actual" differences found in the data between the two types of programs are still reported. Further research and/or the application of more effective statistical procedures will be needed before the reader should view this portion of the study as other than "interesting" (as opposed to "conclusive").

Use

The findings in this category (with one exception) are what might be expected. State-oriented programs more often use their assessment results for setting priorities under Title III of ESEA, use the results to measure state adopted goals, and use the results to influence allocation of state or federal funds. Locally-oriented programs, on the

other hand, are more frequently designed so the results can be used to compare one district against another. No difference was observed between the two groups in the use of programs designed to provide both "predictive" and "actual" scores for districts. This last finding is somewhat surprising since this type of approach is frequently justified as producing data of specific use to participating districts.

Authority

This dimension also developed as expected. State-oriented programs more frequently require local district participation and are more frequently established by legislative mandate (21% as opposed to 7%). As a comment on this factor, the author was told by several assessment directors that while legislators have sometimes required statewide assessment programs under the guise of improving state-level decision-making, actual operation of the program has revealed such legislators to really be primarily interested in the test results of specific districts. This potential mis-match between public purpose and private interest is another area worthy of future research.

Methodology

As noted earlier, this category revealed the only statistically significant difference between those programs designed primarily to produce data for use by participating districts as opposed to those programs aimed primarily at state-level decision-making. It was found that locally-oriented programs used the standardized, norm-referenced type of measurement much more frequently (86% of the time as compared to only 36% by state-oriented programs). Using Fisher's Z Test of Differences for Uncorrelated Proportions, this correlation turned out to be 3.06 (1.96 is needed at the .05 level of confidence). This result was not too surprising. Given the cost of developing

alternative instruments and the larger number of districts usually included in the locally-oriented programs, it may not be feasible for this group of states to move wholesale to criterion-referenced approaches. However, the national debate over the value of standardized, norm-referenced instrumentation might lead one to expect more than only 2 states with a local orientation to be using a criterion-referenced approach. Research into why states with a local orientation to their assessment program are not moving more rapidly to criterion-referenced instrumentation would be useful.

In terms of population design, 82% of state-oriented programs used a sampling approach, while 72% of locally-oriented programs tested all children in the state in specified grades. Only state-oriented programs used both approaches simultaneously (i.e., sampling in some grade levels and testing all children in others).

Scope

In looking at grades to be tested, it was assumed that since locally-oriented programs overwhelmingly use norm-referenced instruments, the pattern would show such states concentrating on the elementary grades. This assumption was based on the fact that most local schools test with such instruments more frequently in the elementary grades than in high school. Also, proponents of standardized, norm-referenced instruments have argued that there is more national agreement on objectives for basic skills development in the elementary grades than for other grade levels. The assumption was proven invalid. Table V shows the distribution between the two groups for each grade, K-12.

Locally-oriented programs tended more often to conduct testing in the junior high and secondary grades with the exception of the fourth grade which was a preferred testing

TABLE V: GRADES TESTED REPORTED BY STATE V.S. LOCALLY-ORIENTED PROGRAMS *

GRADE	STATE-ORIENTED PROGRAMS (N-28)	LOCALLY-ORIENTED PROGRAMS (N-14)
K	7%	0%
1	11%	21%
2	21%	7%
3	32%	29%
4	68%	43%
5	25%	36%
6	29%	29%
7	32%	29%
8	29%	57%
9	18%	36%
10	7%	21%
11	32%	50%
12	21%	36%

* Since most states conduct tests in more than one grade, the percentages in each column total more than 100%

grade by a majority of all states. Only state-oriented programs conducted assessment of young adults (usually defined as age 18-21), with 5% of such states testing this population.

A different distribution also was generated when the variable of domain to be tested was examined. These results are shown in Table VI. As noted, state-oriented programs were much more likely to test only cognitive areas, while the major pattern in locally-oriented programs was to focus on both cognitive and affective domains.

TABLE VI: TESTING DOMAINS REPORTED BY STATE V.S. LOCALLY-ORIENTED PROGRAMS

DOMAIN	STATE-ORIENTED PROGRAM (N-28)	LOCALLY-ORIENTED PROGRAM (N-14)
COGNITIVE ONLY	46%	29%
AFFECTIVE ONLY	4%	15%
PSYCHOMOTOR ONLY	4%	0%
ALL DOMAINS	7%	14%
COGNITIVE AND AFFECTIVE ONLY	39%	42%

In looking at the affective area, no differences were observed in the two groups in terms of whether they choose to focus on school-related attitudes or self-home-community attitudes.

Little difference in the range of subject matter tested was observed in the two groups. State-oriented groups tested in 16 fields while locally-oriented programs tested in 15. However, some variation did occur when the specific subject matter was examined.

TABLE VII: SUBJECT MATTER* TESTED BY STATE V.S. LOCALLY-ORIENTED PROGRAM

SUBJECT MATTER	STATE-ORIENTED (N-28)	LOCALLY-ORIENTED (N-14)
Mathematics	75%	78%
Reading	89%	64%
Science	36%	57%
Social Studies	25%	43%
Career Education	75%	14%
English	21%	36%
Language Arts	11%	21%
Writing	11%	21%
Aptitude	11%	21%

*Only those subjects in which at least 10% of the states initiated testing are included.

As shown in Table VII, state-oriented programs showed a far greater emphasis on the field of reading than did locally-oriented programs. It may be the latter feel their local districts already have access to reading data through district testing programs. On the other hand, it may be that state-oriented programs have been more sensitive to the public pressure to determine "whether Johnny can read." Locally-oriented programs generally showed a wider span of interest in their selection of field of testing, indicating that perhaps this type of program is used by participating schools for the purpose of curricula-wide diagnosis.

The spread of subject matter areas covered by locally-oriented programs may also be a function of their more frequent selection of commercially-produced norm-referenced instruments in that they choose to use all of the subtests available from such sources.

Summary

With few exceptions, this section of the study suggested that there may be differences on the analysis questions between those states who orient their program towards state-level decision-making and those intending to produce information for local school use. The availability of conclusive findings was hampered by the small sample involved. In almost all cases, these tentative differences were "true" to the particular purpose indicated; that is, the differences recorded for state-oriented programs were those which one would expect if the program were truly aimed at state-level activities. In this regard, it can be assumed that states have tried rather conscientiously to select the methodology, population design, scope of testing and so forth best suited to the overall purpose of their assessment program. Findings included:

- State-level-aimed programs tend to use their data for setting Title III ESEA priorities, for determining progress towards state goals and for influencing the allocation of state or federal funds.
- Locally-oriented programs were more often designed to allow for comparison between districts.
- State-oriented programs were more frequently mandated by the legislature and more often require local district participation.
- Programs designed to produce information for statewide decision-making rely most frequently (57%) on criterion-referenced models of measurement.
- Locally-oriented programs almost always (86%) use the standardized, norm-referenced type of measurement and only 2% of these states use any form of objective-referenced instrument.*
- The overwhelming majority (86%) of state-aimed programs rely on a sampling approach, while 72% of all locally-oriented programs test all children in the state in specified grades.
- Locally-oriented programs do not concentrate their testing in the elementary grades where standardized norm-referenced instruments are more usually employed. Instead, they tend to test (in rank order) grades 8, 11, 4, 9 and 5.

---

* This was the only finding where the difference between the two types of programs was statistically significant at the .05 level.

- State-aimed programs, on the other hand, test most frequently in grade 4 (86%), with grades 11, 7 and 3 running a distant second.

- State-oriented programs were also more likely to test only in the cognitive domain (46%), while programs producing information primarily for local schools preferred to assess both cognitive and affective areas (42%).

- State-oriented programs concentrate most heavily on the assessment of reading (89%), while only 64% of the locally-aimed programs test in this field.

- Locally-oriented programs, on the other hand, show a much wider span of interest in selecting their curricula areas, with over 40% of these states assessing in mathematics, reading, science, and social studies.

Part III: Strengths and Weaknesses of Four Types of Measurement

A re-examination of the state responses when grouped by the type of measurement employed offered some clues as to the apparent strengths or weaknesses of each approach. A review of the states' own assessment material also produced additional insights, as did discussions during the phone interviews.

In general, however, these findings must be labeled as "tentative" and "inconclusive." The major difficulty in securing significant results in this dimension was the states' own inability to specify exactly what types of decisions they wished to make as a result of their assessment effort. Measurement modes, of themselves, are not inherently "good" or "bad." They can only be judged so if they are or are not producing the type of information desired by the program administrators and audiences. Until these parties are more precise about exactly what information they want, only limited and subjective analysis can be expected in this field.

Standardized: As noted before, this type of measurement was most likely to be utilized by assessment programs reportedly producing information for local district use. A series of strengths were thus found, as follows:

- The instruments (or type of instrument) are familiar to a large number of local school personnel. These personnel already have experience in both administering and interpreting this mode of testing and conceivably, are thus more likely to know the dimensions of decisions which can be made with the test results.

- Large numbers of children can be assessed for relatively low and known cost. Remember, these states are more likely to be testing all children in the state at specified grade levels.

- A wide range of technically validated and reliable instruments are available in a span of curricula areas. Schools wanting data from the state program do not need to wait until special instruments are developed.

- States reported the ability to reach rather easy and rapid decisions on the choice of instruments. Since objectives were already predetermined by the test manufacturers, considerable time and effort was not needed to try to achieve statewide agreement on what was or should be the major focus of schooling curricula.

Several weaknesses were also noted:

- States found it more difficult to initiate assessments in emerging priority areas such as citizenship or to assess skills on an inter-disciplinary basis. Since the production of testing devices is an expensive undertaking, commercial test publishers are understandably hesitant about preparing standardized, nationally-normed instruments which may be used in only one or two states.

- States reported that there was increasing dialogue as to whether such instruments were, in fact, focusing on the highest priority skills and knowledges. The growing debate over whether norm-referenced tests do not ultimately lead to the rejection of items which "every student should know," is of particular concern to those states still solely dependent on such modes of measurement. At least five of the states now using the norm-referenced model reported that they were either discussing or in fact actually planning to begin development of criterion-referenced instruments.

- Some states also indicated that they felt an essential outcome from statewide assessment - the ability to get large numbers of people to discuss what should be the purposes of schooling - had been lost when they decided to rely on the norm-referenced model. They noted that since the design of objectives was not required by this approach, their program had not contributed much to a clarification of the public's expectations of schooling.

National Assessment Model: A surprising result of the study was the documentation of the degree to which the National Assessment of Educational Progress has influenced the assessment programs of individual states. While only six states were actually relying primarily on NAEP objectives and test instruments, it was found that an additional ten states reported using some of the objectives or test items in their own

criterion-referenced program development. Eight more states reported using either NAEP administration procedures, sampling plan, measurement materials, or dissemination documents as models for activities in their own program. Strengths reported for this model were as follows:

- The program has produced criterion-referenced instruments in areas where individual states say they would not have had the resources to develop their own, such as citizenship and music.
- The model has made instruments available for populations not easily measured through the standardized, norm-referenced approach, i.e., 17 year olds and "young adults."
- The model has been useful to states who wanted to be able to compare their state or district results against national results, but were also committed to the use of criterion-referenced instrumentation.

Three kinds of criticisms or weaknesses were reported for this model, however.

- NAEP releases only half of its objectives and test items in a subject matter for state use. While an understandable policy for the national program, it limits the coverage which can be achieved by an individual state.
- States which rely primarily on NAEP material must necessarily use NAEP objectives, and again these may not cover areas thought to be priorities in a particular state. Some states report overcoming this difficulty by adding a small percentage (usually no more than 25%) of their own objectives to those received from NAEP. States have not reported finding a method for scheduling their assessment activities at times and in years other than those specified by NAEP unless they do not wish to make national comparisons.
- States using the NAEP objectives and test items are also dependent on the overall quality of the national program. When a poor test is developed in a particular field (as has been charged in the case of NAEP's recent social studies assessment), the state has no alternative available. In addition, cuts in the NAEP budget or staff (as has been witnessed in 1973-74) also limits the states' programs.

Objective-Referenced: An increasing number of states appear to be moving towards the development of their objective-referenced measurement model. In most cases

these tests are still in the formative or experimental stage and it is too early to make definitive evaluations on the effectiveness of this approach. However, some comments can be made at this time.

- The major strength of this model appears to be the states' ability to decide for itself what are the critical skills, knowledges and attitudes for their locale. States thus report that they expect the consumers of the program data to have far greater confidence and interest in the assessment results.

- The development of such models has been found by states to be an effective vehicle for launching statewide dialogue on what should be the expected outcomes of schooling. Since statewide assessment programs were initiated primarily as a methodology for achieving "accountability," this is seen as a definite plus.

- States report that they also see this model as leading to a more humane and responsible attitude towards education, i.e., to focus on what every pupil should (or does) know, as opposed to focusing primarily on areas where there are differences in student performance.

Some problems have been encountered, however, which must be classified tentatively as "weaknesses."

- The development of criterion-referenced instruments tailored to a single state's needs is a costly and time-consuming enterprise. Adopting this model usually means that states have been able to assess in only one or two subject matter areas a year. They report growing concern that the financers of assessment and the potential audience will become discouraged with both the cost and the slowness with which a comprehensive, multi-subject-matter program can be developed.

- Questions of the reliability and validity of state-developed instruments are already being raised. While research on this dimension is in its infancy, there is some evidence that this may become a major problem.(26)

- States moving to adoption of this model report the need for an extensive "education" program for both professional school personnel and the general public to insure that results are disseminated and understood correctly.

- Methodology required to implement some aspects of this model are also lacking and few states expect to have the R & D funds necessary to fill these gaps. Of particular concern is the development of instrumentation in the affective domain, the identification of methods needed to assess objectives in an "applied" or simulated setting and the understanding of the most effective way of identifying "mastery" objectives and standards.

Eclectic: This measurement model was found to be used by only a limited number of states and it is not an approach that is being widely advocated. For these reasons, the review will be succinct.

- The major strength of this model is its reliance on instruments already being administered for other purposes in the state. For this reason, the model may be viewed more as a way of "gathering and reporting" data than a model for actually generating its own data.
- The model is seen as a relatively inexpensive method for initiating statewide assessment when funds do not permit the purchase or development of new measurement modes. For this reason, the model also maximizes the use of data generated primarily for other purposes.

Several weaknesses are obvious.

- The content and coverage of this type of measurement are totally dependent on the availability of data from other activities in the state. It is difficult, therefore, to design any comprehensive assessment effort or to decide on measuring high priority skills, knowledges or attitudes of interest to the state but not included in ongoing testing efforts. One state using the eclectic model has attempted to deal with this problem by administering one standardized, norm-referenced test in a single subject matter each year as a way to supplement data available from other sources.
- The variety of testing devices used, the diversity of populations included and the variation in subjects makes it difficult to develop comprehensive statewide assessment reports. One state using this model said they called their annual report a "fruit salad," because of the necessity for blending data of such variation. They also reported that for this reason, they doubted whether the assessment data had much impact on either state or local decision-makers.

Summary

In conclusion, each of the four types of measurement have been found to have both strengths and weaknesses. More extensive evaluation on the effectiveness of the various approaches will be dependent on greater clarity of states about the relationship between their statewide assessment program and desired decision-making.

CHAPTER IV: RECOMMENDATIONS FOR FUTURE RESEARCH

This paper has been written in part to provide both school practitioners and educational researchers with a "beginning point" for determining future actions related to statewide assessment programs. It has attempted to identify and classify the 42 operating programs along such dimensions as purpose and use, authority, methodology and scope. Throughout the paper, comments have been made about the inability to draw definite conclusions or to answer interesting questions because of the unavailability of research in this field. As noted in Chapter II, with the exception of a few R & D efforts operated primarily by the states themselves (or consortia of states),only spotty attention has thus far been paid to this exciting phenomena by the professional research community.

An exhaustive listing of all of the research needed in the field would constitute a paper in its own right. But this chapter attempts to highlight some of the more significant needs.

- Immediate research is needed on the question of the most effective roles for statewide assessment programs in influencing state or local decision-making. Is it appropriate for statewide programs to be producing information primarily for the use of local school districts? If so, how does this information differ from data which would be generated by the districts' own testing program? What types of decisions should state governing boards or legislatures make as a result of state assessment? Should this type of information be used in allocating funds to state programs or to local school districts? If so, should the state reward those who are doing well or provide additional monies for those are are doing poorly?

- Given that roles of statewide assessment programs are defined, how does a state insure that assessment data is actually used in making the desired decisions? Is it appropriate to expect that public policymaking and/or budget decisions can be shifted to a data basis (as opposed to political factors, personal preferences or other basis for decision-making). If so, exactly how is this to occur? What are the most effective means for disseminating assessment results to target audiences? What type of follow-up activities are most effective in getting desired results?

- Research is also needed on procedures and techniques which will lead to wider availability of criterion-referenced instruments. Increased options for determining the technical reliability and validity of such instruments are needed. Better techniques for establishing both valid and politically defensible standards of performance for interpreting results from such tests are required. How can the currently high costs associated with moving to this form of testing be reduced? Can options be increased for those states who want to develop their own criterion-referenced instruments, but whose legislatures are requiring comparative data with other states?

- Research is also needed on a whole range of philosophical questions underlying the move towards criterion-referenced measurement. Should objectives be aimed at measuring what most students can do or what adults think they should be able to do? Are the skills and knowledges traditionally used as the focus for such objectives really the essential things to be measured? Should states not instead be looking for other types of abilities, i.e., flexibility, adaptability, tolerance for change, ability to define problems and so forth?

- Studies are needed immediately which will solve some of the methodological problems facing state assessment programs. What are effective measurement modes which can serve as alternatives to the traditional reliance on paper and pencil tests? How can states measure competencies which they want to be demonstrated in a "real" (or near real) setting? What are the most effective means for measuring attitudes, beliefs, values? What types of experimental designs might be appropriately applied to state assessment programs? What student, school and community variables appear to correlate most significantly with student achievement and how can these be measured most effectively?

- Finally, research of an evaluative nature on the whole phenomena of statewide assessment should be undertaken. The U.S. Office of Education estimates that at least $5 million annually is now being spent by State Departments of Education to produce assessment data of the type required by Title III, ESEA (27). Is this expenditure really producing any tangible benefits? Are better decisions really being made? Given Leon Lessinger's "three basic rights of education" cited at the beginning of this paper, has the phenomena of statewide assessment really proven an answer to the question of how to secure educational accountability? Are there more effective options which are being overlooked in the rapid movement of all states to the assessment "bandwagon?" Would states be better off spending their current assessment budget on helping all local school districts develop their own data-based management systems?

Lastly, it is the recommendation of the author that some federal agency be given the responsibility (and funds) to conduct or contract for an annual survey of the status of statewide assessment programs and to disseminate these results widely. Not only

would such a survey continue to stimulate additional research needs, but it would insure that states had an opportunity to keep abreast of new developments and possibly avoid the same costly mistakes. The work of the ETS Center for Statewide Educational Assessment, the State Education Accountability Repository and the Cooperative Accountability Project have all provided a valuable beginning in recent years. However, the project-based funding of all three centers is now reaching a close and unless such activities become part of the ongoing responsibility of a national agency or organization, this critical source of information will end.

FOOTNOTES

(1) Brademas, John, "Accountability: A Rationale", speech delivered at a symposium on accountability in education sponsored by Memphis State University, March 1973. Excerpts printed in Commentary: Cooperative Accountability Project, Vol. 1, No. 1, Colorado State Department of Education, Denver, Colorado, March 1974, p.1.

(2) Fortna, Richard O. (ed.), State Educational Assessment Programs: 1973 Revision, Center for Statewide Educational Assessment and ERIC Clearinghouse on Tests, Measurement and Evaluation, Educational Testing Service, Princeton, New Jersey, 1973, p. 1.

(3) House, Ernest R., Rivers, Wendell, and Stufflebeam, Daniel L., "An Assessment of the Michigan Accountability System," Phi Delta Kappan, Vol. LV, No. 10, Bloomington, Indiana, June, 1974, p. 664.

(4) Womer, Frank B., Developing a Large Scale Assessment Program, Cooperative Accountability Project, Denver, Colorado, 1973, p. 7.

(5) Beymer, Lawrence, "The Pros and Cons of the National Assessment Project," Contemporary Issues in Educational Psychology, Clarizio, Harvey F., et. al. (ed.), Allym and Bacon, Inc., Boston, Mass. 1970, pp. 426-431.

(6) Finley, Carmen J. and Berdie, Frances S., The National Assessment Approach to Exercise Development, National Assessment of Educational Progress, Ann Arbor, Michigan, 1970.

(7) Conaway, Larry E., "Some Implications of the National Assessment Model as Data for State and Local Education," paper presented at the 58th Annual Meeting of American Educational Research Association, New Orleans, Louisiana, February, 1973.

(8) Buchmiller, Archie A., "State Assessment: Potential for Becoming a Friend or Foe?" Speech delivered at 78th Annual Meeting of the North Central Association of Colleges and Secondary Schools, Chicago, Illinois, March 1973.

(9) Hawthorne, Phyllis (ed.), Legislation by the States: Accountability and Assessment in Education, Cooperative Accountability Project, Denver, Colorado. Issues available for 1972, 1973 and 1974.

(10) House, Rivers and Stufflebeam, p. 664.

(11) Fortna, p. 1.

(12) Campbell, Paul B. The Use of Correlates of Achievement in Statewide Assessment, Center for Statewide Educational Assessment, Educational Testing Service, Princeton, New Jersey, 1973.

(13) The University of the State of New York, Variables Related to Student Performance and Resource Allocation Decisions at the School District Level, The State Department of Education, Albany, New York, June 1972.

(14) Hall, Mary, "Summary of State Assessment Models," Statewide Assessment for Oregon Schools, Randi Douglas (ed.), College of Education, University of Oregon, February 1973, pp. 79-93.

(15) Fortna, p. 1.

(16) Joselyn, Gary, et. al., State Testing Programs; 1973 Revision, Educational Testing Service, Princeton, New Jersey, 1973.

(17) Fortna, p. 1.

(18) Joselyn, 1973.

(19) House, Rivers and Stufflebeam, p. 664.

(20) Fortna, p. 1.

(21) House, Rivers and Stufflebeam, p. 664.

(22) House, Ernest, et. al., "An Assessment of the Michigan Accountability System," p. 668.

(23) Fortna, p.1.

(24) The University of the State of New York, 1972.

(25) Fortna, p. 1.

(26) House, Rivers and Stufflebeam, p. 664.

(27) Hershkowitz, Martin, Statewide Educational Needs Assessment: Results from Selected Model States, Hershkowitz Associates, Silver Springs, Maryland, 1974, p.7.