DOCUMENT RESUME

ED 095 202 80 TM 003 874

AUTHOR Tronsgard, David T.; And Others
TITLE Statewide Educational Evaluation.
INSTITUTION National Association of State Boards of Education, Denver, Colo.
SPONS AGENCY Bureau of Elementary and Secondary Education (DHEW/OE), Washington, D.C. Div. of State Agency Cooperation.
PUB DATE 74
NOTE 78p.
AVAILABLE FROM National Association of State Boards of Education, 2480 W. 26th Avenue, Suite 215-B, Denver, Colcrado 80211 ($3.00)

EDRS PRICE MF-$0.75 HC-$4.20 PLUS POSTAGE
DESCRIPTORS Academic Achievement; *Educational Assessment; Evaluation Criteria; *Evaluation Methods; Information Dissemination; Learning; *State Boards of Education; *State Programs; Statistics; Testing; *Testing Programs
IDENTIFIERS Elementary Secondary Education Act Title V; ESEA Title V

ABSTRACT

This book has been written for state board of education members and other citizens interested in public education. It is, in a sense, a primer in matters relating to learning, testing, assessment, and evaluation. Presented are some philosophical and political considerations in statewide educational evaluation. Learning is defined and the types and levels of learning are discussed. The remaining sections are devoted to: the measurement and evaluation of studnet learning; the problem of appropriate educational criteria; some suggestions for reporting the outcomes of evaluating student learning; and some rules of thumb which state school board members might employ to assist them in the evaluation of student learning resulting from curricula under their purview. The appendixes contain a directory of key state educational evaluation personnel and contracting agencies used by state education departments for matters relating to assessment. (Author/RC)

STATEWIDE EDUCATIONAL EVALUATION

David T. Tronsgard
Michael J. Grady, Jr.
E. Dean Coon

The National Association of State Boards of Education
Denver, Colorado
1974

Financed by Funds provided by the Elementary and Secondary Education Act of 1965 (Public Law 89-10, Title V, Section 505) and greatly assisted by the sponsoring and cooperating states:

New York
Alabama
Nevada
New Jersey
Kansas
Colorado

Copies of this book may be obtained from:
National Association of State Boards of Education,
2480 West 26th Avenue - Suite 215-B
Denver, Colorado 80211

FOREWORD

This little book has been written for state board of education members and other citizens interested in public education. It is, in a sense, a primer in matters relating to learning, testing, assessment and evaluation. One unusual feature of the book is that it begins with a glossary. Educators are famous for their argot. Often they use complicated words for simple ideas. In the case of words relating to evaluation, the technical words are usually more justified and the meanings precise. For this reason, the authors of the book believe that laymen will want to learn the most common meanings for the jargon relating to learning and measurement. These words are used in the text and hopefully as defined.

Reading the book should prove as helpful exercise prior to important discussions of statewide assessment. The issues delineated should at least be considered before decisions are made. After all, very important educational and social consequences are at stake.

David T. Tronsgard
Executive Secretary
NASBE

ABOUT THE AUTHORS

David T. Tronsgard has served for four years as Executive Secretary of NASBE. Prior to that he was a college professor and administrator, and a superintendent of schools. He has an A.B. degree from San Diego State College, an M.A. degree from Teachers College, Columbia University and an Ed.D. degree from Stanford University.

Michael J. Grady, Jr., is presently an independent educational consultant. A retired Air Force Lieutenant Colonel, Mr. Grady has previously been a director of research and an employee of the Colorado Department of Education. He received an Ed.B. degree from Rhode Island College, and M.A. and Ph.D. degrees from the University of Alabama, the latter in educational and counseling psychology.

E. Dean Coon is currently an associate professor and assistant director for the Center for Northern Educational Research, University of Alaska. Previously he was Associate Commissioner of Education in Colorado. He received his doctorate from the University of Denver. His M.A. degree is from University of Nebraska and his A.B. degree from Colorado State College in Greeley.

TABLE OF CONTENTS

GLOSSARY

Accountability - The process of being accountable. In education, accountability refers to the proper stewardship of educational funds, resources, and the generation of appropriate service for learners. It may be defined narrowly, as by legislative mandate for public disclosure, or broadly, as by parental expectations.

Affective learning - Learning which describes changes in interest, attitudes, and values, and the development of appreciations and adequate adjustment. Citizenship skills, for instance, are affective.

Arithmetic mean - Average of all the scores in the distribution: the scores added and divided by the total number of scores.

Cognitive learning - Learning which deals with the recall and recognition of knowledge and the development of intellectual abilities and skills; academic learning.

Comprehensiveness - A test is comprehensive if the test items adequately sample the full range of objectives and subject matter in a curriculum; the degree to which a test is narrow or broad.

Criterion-referenced test - A test designed to measure the achievement or mastery of school objectives; a test reflecting pre-determined values, judgment, or professional standards and scored *a priori* without regard to norms.

Differentiation - In testing, when test items discriminate between bright students who get the items correct and slow students who get the items wrong.

Educational deprivation - Educational practices or cultural exigencies which do not allow students to learn in accordance with their abilities.

Effect (learning law of effect) - Man learns what is pleasurable or is satisfying to learn.

Evaluation	-	Judgment based upon criteria; opinions and values applied to assessment.
Explication	-	A process of systematically describing, or the process of dividing a topic into semantic or logical components.
Frequency (learning law of frequency)	-	Things learned most often are best remembered.
Goal	-	Broad statement of educational purpose.
Hypothesis	-	A tentative statement of an educational outcome which a researcher wants to test; a theory to be subjected to investigation.
Imply	-	Communicating a conclusion to someone else, suggesting an interpretation.
Infer	-	Drawing conclusions from data, usually for oneself; coming to a conclusion. One does not infer to or for someone else.
Learning	-	A change in behavior modified by experience. Scholars doubt learning not accompanied by a change in behavior. To say *learning* is the mastery of material gives many semantic problems.
Learning style	-	Specific approach which each learner finds most advantageous to his learning (Visual, auditory, or combination).
Median	-	The score which is half-way in the distribution so that 50% of the scores fall above and below. The median is usually very stable and is not affected by a few extreme scores.
Mode	-	The most frequent score that occurred in a given distribution. In a bimodal distribution, a curve will have two modes (peaks).
Multiple reinforcement	-	Systematic repeating of what is to be learned and the rewarding of correct responses.
Norm-referenced tests	-	Tests which are standardized against the performance of a national sample. Tests which are reported in standard scores based upon the normal curve.

Objective (learning objective)	-	A limited amount of student learning to partially achieve an educational goal; a desired end of a lesson or course.
Objectivity	-	A test has objectivity if the same score is assigned no matter who corrects the test. Essay tests are low in objectivity; multiple-choice tests high.
Primacy (learning law of primacy)	-	Things learned first are best remembered.
Psycho-motor	-	Learning which deals with manipulative or motor-skills.
Recency (law of recency)	-	Things learned last are best remembered.
Reliability	-	Consistency of test results from sequential administrations; the degree to which a test will give the same results when administered again.
Serendipity	-	The process of discovering something important or beneficial by sheer accident, usually when investigating something else.
Sociogram	-	A technique which measures likes and dislikes of members of a group in regard to other members of a group. A diagram of group preferences within a group.
Statewide assessment	-	Gathering data about educational learning on a statewide basis. Measuring the status of educational achievement.
Taxonomy	-	An organized listing, the end result of an explication; the classification of phenomena.
Validity	-	The degree to which a test measures what it was intended to measure and nothing else. Proof the test gives the desired information.

CHAPTER I

STATEWIDE EDUCATIONAL EVALUATION

Some Philosophical Considerations

It is fashionable for citizens and their representatives to say they want to measure what the schools are doing, to ascertain what the students are learning, and to see if educational monies are being well spent. What is more, people are beginning to want to compare schools, school districts, counties and even states. Because these facts are so, projects for statewide assessment have begun in a large number of states. National studies and commissions have been organized, and several private corporations have *re-tooled* to aid in these endeavors.

Certainly, it is true, our society lacks a systematic method for appraising the quality of its schools. We have at our disposal many means for assessing student performance, especially by the *norm-reference* method. We can muster impressive data about individuals. However, little of this is systematic and neither can this information tell us much about the quality of schools and school districts. We need, it would seem, better ways of looking at what we are doing, perhaps to include comparisons with other societies.

Testing and publishing companies, even those which are chartered as non-profit organizations, certainly can be expected to react to new demands. They must compete for a share of educational funds so vital to their continuance. This greatly explains their willingness to cooperate in statewide assessment projects. Some politicians and elected representatives of the people have been sensitive to popular demands for evaluation and assessment. Services deemed necessary by the majority (or the credibly noisy) will eventually be implemented. Public pressure explains much willingness to participate in and to allocate resources for assessment. Explanations of the motives behind the rather general clamor among the public for assessment are harder and require more inspection, more introspection, and less cynicism.

Perhaps it could be argued, though not convincingly, the press for assessment was sold to the people by their leaders and by the professional testors through the mass media, and now these same persons are merely responding to the induced demands. In any case, the demand is present and must be served. How to serve it wisely is the problem. This handbook is meant to help state board of education members consider the philosophical, political and technical ramifications of the measurement of school achievement. It is also intended to be a resource book for all citizens who want to study both the rewards and dangers inherent in statewide assessment of educational efforts.

If we totally assess the full spectrum of educational accomplishment and know how all our students are performing in all their subjects, and know likewise how our schools stand in relation to each other, and how our school districts are doing, what will we have proved? Before we start such a major task, we must certainly realize that no matter how good we are, one half of all our pupils will score average or below in relation to the total pupil population; one half of our schools and school districts will do likewise. Also we will find that middle and upper-class students will in general get *more than their share* of the better marks. Why then should state board of education members be asked to support a movement which will *discover* what we in general already know?

Starting with a *brain-set* that assessment is best accomplished by norm-referenced measurement is probably a lack of open-mindedness. Good statewide assessment may make only limited use of norms and probably will focus instead on criteria. These criteria may be logical, predetermined, and arbitrary. Practical accomplishment of goals is what we are after and assessment can tell us if we are succeeding.

If we can generate an educational structure which both will allow the vast majority to become self-reliant and productive adults and will have flexibility to serve individuals throughout their lives, differential performance may prove useful rather than discouraging. Bearing in mind that statewide assessment is analysis of our system, as much as it is measurement of individuals, we should not be led astray by discrepancies in individual potentialities. To justify assessment we need reasons for it and to paraphrase a previous point, state board members should be able to recite at least a few.

The practical comes to mind first. State boards must make recommendations to legislatures. They often determine minimum curricular goals. They must allocate resources, at least those of state departments of education. Surely they help set state priorities for education and participate in planning. Assessment can provide the raw material for these kinds of choices as well as the test of them. This is a good reason for assessment. Others are more subtle but equally valid.

Some of us are curious, a large number of us. The reasons are many and a few of them should be explored here. We want to measure school performance because we do believe it to be measureable and we really want to see what people know, can know, and ought to know. Like pure scientists, some of us adopt the optimistic hypothesis that we will discover good uses for the information if we only had it, just as many uses for the population census have been discovered beyond the imagination of our founding fathers. In the past we have discovered that test results are useful for individual counseling, are

useful for discovering correctable weaknesses, and can aid in curriculum construction. The goals of past testing programs were primarily directed at individual uses; although school administrators fortunate enough to head districts of high scoring pupils may have used test results for what could be called *propaganda* purposes. Faith in serendipity may not be a good argument to use in advocating statewide assessment to taxpayers, especially considering the inevitable high costs. But honesty requires this admission by many.

Another reason some of us want statewide assessment is in order to make comparisons possible. We can see if districts which spend more money do better. We can see which of our school citizens have high achievement and where we are getting the most for our money. At the same time we can find our weaknesses so that we can admit them, and hopefully correct them. We are not sure of the full consequences of comparisons; neither have we deduced a complete explication from among the infinity of possible comparisons. Honesty again requires us to admit that we know all school districts are spending to or near the limit of their incomes and can always make use of more funds. In other words, rich districts almost always spend more money per pupil. We also know that our big cities will not do as well academically as their suburbs regardless of edueational expenditures.

Accountability is a fourth reason for statewide assessment, one which has much popular appeal and derives from the ability to compare results. Some of us believe that if we can make comparisons, we can hold schools and school districts accountable. We can say they have or have not been faithful stewards of our children and our resources. Much of the steam behind demands for statewide assessment obviously stems from the hope that accountability is both feasible and possible. Prevalent, of course, among *taxpayers* is the belief that many educational professionals are *voracious rascals* creative only in their collective ability to squander public monies. Accountability is seen as one way to curb these appetites. Some demands for accountability are based upon the assumption that assessment can prove how good individual teachers are. The consideration, to use a most apt cliche, is *a real can of worms.* It nevertheless exists as a motive for assessment.

A fifth push for statewide assessment comes from advocates of the view that education is partly responsible for the amelioration of differences among children due to economic, intellectual, or physical reasons. Because we espouse compassionate ideals of equity, many of us want to treat the unfortunate unequally to give them a favorable chance to compete in our highly competitive society. Inherent in these considerations are many philosophical contradictions (regardless of one's point of view), but the argument to augment educational op-

portunities for the poor, the handicapped, and the educationally deprived persists. Educational deprivation can be demonstrated by testing, measurement and assessment. It proves what would otherwise be an assumption. It follows then that areas of poor educational performance can be attacked and reduced. In an educational system based upon arbitrary standards, this is a valid argument. Measurement based upon the relativity of the normal curve of distribution will show the same number of *winners* and *losers*. However, the mean for any curve can be raised. A pessimist would say that wanting to reduce inequalities is easier than the finding of solutions to it. Certain recent publications add to pessimism (especially Jencks, *Inequality*) but most of us prefer optimism and are sure we can make things better.

The last reason given here for statewide assessment is that it will aid us in decision making. Decision making stems from and relates to the practical. We will know where we are strong and where we are weak. We will also know who is strong and who is weak in an academic sense. Armed with this information, we are in a better position to make intelligent decisions. State board of education members make many decisions, most of them on the basis of personal opinions and values, and some of them because of public pressure. With the results of statewide assessment in their hands, state board members can make decisions more scientifically and more correctly. They can also answer the criticisms that arise about decisions more precisely.

To summarize the six reasons given here for statewide assessment they are as follows:

1. practicality
2. curiosity
3. comparison
4. accountability
5. amelioration
6. decision making

Certain articles of faith accompany these motives to assess. The first is the universally held assumption that *the way things are done in the schools will make a difference in the academic performance of individuals*. If we didn't believe this, all our efforts would seem futile. Much of the research on curricula, probably most of it, hasn't been encouraging. We have made only a few breakthroughs in procedures which dramatically change how children learn but we must always assume more are possible. What alternatives do we have to an assumption that changes in allocations, particularly financial allocations, will change achievement patterns? Recently critics have marshalled, we must admit, impressive evidence apparently debunking the theory that increased educational expenditures improve educational performance. These critics have at least shown that unexamined acceptance

of this article of faith is at least naive and possibly false. Thirdly, assessing educational accomplishments assumes that we know what causes learning: if we know how students do, we also know why they do thusly. This means we are willing to grade the quality of our inputs while we are grading the efforts of the students. A fourth article of faith is based on the belief that we can cause and control change; equality is possible; amelioration is feasible; we can do something about areas of weakness. If we do not believe at least one of the foregoing articles of faith, then we should not waste much time on assessment. We should merely follow traditions, public wishes, and professional recommendations and worry less.

Before we commence any widespread efforts to assess what the schools are doing, we have an obligation to know what we are doing in education and why. As decision makers in the educational process, state board members should be prepared with answers with what they believe desirable in education. Their values and philosophical ideas should be rational and well-defined. Examples of questions educational decision makers should be able to answer are contained in the following *taxonomy:*

Questions for Educational Decision Makers

What should our school graduates

1. *know?*
2. *be able to do?*
3. *believe?*

To what extent should the schools

4. *ameliorate social class differences?*
5. *emphasize individual excellence?*
6. *insist on common minimum standards?*
7. *grant diplomas and promotion on standards of performance?*
8. *treat the exceptional student exceptionally?*

What do our citizens believe.

9. *should be the goals of education?*
10. *are the essential priorities of education?*
11. *about the value of student preferences and needs?*

What should we do about the contradictions affecting education.

12. *between freedom, progress, and necessity?*
13. *between goals of equality and goals of excellence?*
14. *between what we believe and what we are doing?*
15. *between societal, and individual needs and preferences?*

These are hard questions but when we agree to be educational decision makers, we imply to others that we have answers to these questions or are at least willing to gain them. Deciding what the schools will be like and what they will teach is probably also deciding partly how all other institutions will be. It is morally wrong to take on such a responsibility without the honesty to also attempt to be a

philosopher on matters relating to education. It is true, then, to restate, that major educational decisions are philosophical decisions. The test of decisions is utility, which may be demonstrated by assessment, but the soul of decisions is wisdom and insight. These come from study, debate, and contemplation.

The data collected from statewide assessment has no meaning, as does any data, without interpretation. Interpretation requires frames of reference. Answers to the questions given as examples in the taxonomy of questions can be the reference points. As you noticed the questions were divided into four parts. The first asked about students and their performance. Obviously we have to have something in mind for students before testing makes sense. After we have decided what is good, we can see if our ideals are being approached.

The second group asked questions about social policy. These are questions drawn from the classical arguments relating to treatment of students. In American society we have responded to our educational needs by providing mass education and have somewhat obscured goals for individual performance. We appreciate excellence but we have been more concerned with typical accomplishment. Many of us are now beginning to question our system for this reason. Education by *behavioral objectives* and diplomas for performance are recent examples of proposals to return to a more *European* educational system, and one less egalitarian, therefore. If we use assessment to usher in arbitrary standards, this is not necessarily wrong but we should do it knowingly, anticipating the fallout benefits and detriments.

The third group of questions requires decision makers to know their constituencies. Because we are a democratic republic our decision makers have an obligation to take into consideration *vox populis*. They should not be blind-sided by their own biases, both a dangerous and arrogant attitude.

The fourth group of questions are *gut-level* philosophical issues. These questions show the many paradoxes hiding in what we believe and what we must decide. We do, in fact, believe and act upon many conflicting goals. We make many compromises with ourselves as well as with others. The best stance to assume in the face of contradiction is humility and willingness to study, to listen, and to learn.

In summary, this chapter is meant to give state board of education members a primer in educational philosophy, especially in those aspects relating to problems of measurement of progress toward education expectations. Statewide assessment of education is very costly both in money and time. To justify these expenditures, it seems imperative we know what we are doing and why. Besides just telling us how well we *stack up*, statewide assessment may alter many things including goals and public attitudes. Many will assume assessment to be the end of education itself. We will not want this to happen but it will. Others will use assessment as a club, parsimonious legislators for instance. Others will fear it, perhaps teachers who stand to be

judged by student performance. We don't want to release a box full of evil spirits and pestilence into the world as did the mythical Pandora. We want to make intelligent and rational use of good educational tools and do so with intellectual honesty and fierce personal commitment. This is the spirit in which this book is given to you. The aim is to be utilitarian.

CHAPTER II

STATEWIDE EDUCATIONAL EVALUATION

Some Political Considerations

If after being introduced to statewide educational assessment, your comment is *That must be wonderful; I don't understand it at all*, then you are in good company. Moliere's 17th Century remark is very appropriate today. It must express the feelings of many as they are introduced to the problems and promises of statewide educational assessment.

The previous chapter discussed the philosophy of assessment and suggested some major reasons for such programs. The chapters following will present some technical aspects of statewide educational assessment. The purpose for this chapter is to explore the range of choices open to state boards with regard to assessment, and to identify some of the implications of these choices. Also presented will be a number of the effects which may occur as a result of initiating or expanding a statewide educational assessment program.

Many but not all implications will be presented. It will remain for state board members of the individual states to analyze their own situations, and develop their own lists of: *Here is what will happen if we . . .* This kind of action is vital to effective decision-making by the board on any question.

The Range of Choices

At this point consideration must be given to the range of choices open to state boards; even some investigation of choices apparently closed would be wise. Because a statewide educational assessment program can be so complex, because the planning and operations alternatives are so numerous, and because each state board member's concept of the program will vary, the choices suggested below are reasonably limited.

The choices, five in number, merely offer a range of possible actions which may be taken by a state board of education with regard to an assessment program. The choices are to: (1) ignore it, (2) study it, (3) pilot it, (4) start it, and (5) rush it. Let's examine each choice briefly.

Choice 1: Ignore It. This choice will appeal to some who have found that if something is ignored, it may go away. What better way to save time, money, talk, and frustration than to do nothing, or to postpone consideration for a year.

It is not difficult to find evidence to support this kind of decision. The fourth annual Gallup survey of the public's attitudes toward the schools (*Phi Delta Kappan*, September 1972) reveals that 56 per cent

of the population placed the chief blame for some children doing poorly in school on the children's home life; only six percent blamed the schools and twelve percent blamed the teachers. This logic indicates then that since it isn't the schools' fault that children do poorly, why mount an expensive statewide educational assessment program of limited value?

Choice 2: Study It. This choice is widely used in all kinds of educational endeavors. It's a viable, respected position, since on one hand the problem is not being ignored, and on the other major resources have not yet been committed.

The time allotted and the nature of the study will indicate possible next steps. For example, if there is no time limit set on the study, and the charge is not concise, this choice will be almost the same as *Ignore It.* If, however, a definite time limit is set (ranging from six months to a year), some resources are provided, and a clear charge of what is to be done is given, everyone will know the state board is serious about the potential value of statewide educational assessment. This choice is also useful when an educational issue of more import is at stake: it buys some time. But most importantly, this choice does permit development of a comprehensive plan for statewide educational assessment.

Choice 3: Pilot It. Many advantages are inherent in this approach. It is the choice permitting action short of full commitment. A return to ignoring assessment can be made if the pilot test is aborted or even if it is completed. Piloting a program will probably call for new resources from the districts or the state. So much is going on in assessment that federal and foundation risk money to test and develop programs is short today. Piloting permits legislators, educators, and others to take a look at a *sample* program. Experience gained can be invaluable if a decision is then made to go ahead and develop a program. Legislators anxious for the results, however, will not be content to wait until this phase is completed before having a chance to make the next decision. School year pilot programs tend to be examined at the height of the legislative session, which is usually the middle of the pilot program.

Choice 4: Start It. This is perhaps the most common action that can be taken, especially considering that quite a number of states are already underway with statewide educational assessment programs. It presumes there has been some thought (perhaps even extensive study or piloting) as to the purposes, procedures, resources, and implications. This action indicates that the program will be installed in a state as a regular, ongoing program, subject to continuous evaluation and improvement. It may indicate, *in fact, it should,* that the agencies

controlling education have made a long-term commitment, both philosophical and fiscal, to support the program. It also means that if the long-range planning is not completed, it soon will be, so that an effective, efficient operation will result. Getting an assessment program in this manner, with the limited resource problem, starting an assessment program may mean that some other program may need to be curtailed or abolished.

Choice 5: Rush It. Some will choose rushing, either through design, because of pressures, in order to catch up or get in the lead. Being in the lead, of course, may have current prestige advantages along with some not-so-great consequences. Getting on the assessment bandwagon by this route can lead to some sour notes. Laws mandating a statewide educational assessment program may force a state board to take this choice; the manner and time frame within which the task is accomplished will be important indicators of success. Even counting some of the disadvantages of this choice, a subsequent alternative is to ease back to a more modest program (similar to one which might be operating under *Choice 4*) and retain the best as the program becomes institutionalized. Bundles of money and lots of help will be needed for those states choosing *Rush It.*

Implications Based on Reasons for Assessment

The previous chapter considered six philosophical categories of reasons why states initate statewide educational assessment programs. These were: (1) practicality, (2) curiosity, (3) comparison, (4) accountability, (5) amelioration, and (6) decision making. Let's now consider educational and political elements, and some rather specific motives for assessment, and what the choice of motives might imply about or bring to the program. This list of motives is merely illustrative. The list of implications is not exhaustive.

If this is your motive for a statewide assessment program	then these may be the implications
1. To comply with a legislative mandate	*The legislature is disenchanted with the reporting by the state education agency. *The people and the legislators want the schools to be better. *Better management of scarce resources is imperative.
2. To have better decision-making information at the state level	*Someone has defined the kinds of information needed at the state level. *More state control may be indicated. *State decision-makers will read and use the new reports.

3. To have better decision-making information at the school district level	*This will help maintain local control. *School district information systems do not have common elements.
4. To have better information for reporting the status of the schools to the public.	*The current status of the schools is unknown or poorly defined. *The public has a right to have more and better information. *Knowing the problems and promise of the schools is important.
5. All four of the above.	*This may be an impossible task!
6. To involve more interested persons in the development and evaluation of education programs	*Local control has restricted the flow into the districts of new ideas and outside concerns. *Knowledgeable participants are more likely to support the programs of the schools. *Education shouldn't be left solely to educators.
7. To make state-by-state comparisons of student achievements and attitudes	*Our schools are in need of additional state or federal aid and this information will help justify it. *The state's students are the leaders when compared to others. *The states will use the same assessment instruments.
8. To make district-by-district comparisons	*Competition for excellence will be fostered in the state. *Some local boards and administrators aren't doing a good job. *Districts needing additional help (financial or service) will be identified.
9. To make annual comparisons of the achievement levels of students	*A very meaningful analysis of school programs will be possible. *Teachers will be able to use this information.
10. To obtain documentation for district reorganizations	*Student achievement in small districts may be below average. *Per student costs in small districts are out of sight. *Other reorganization ideas have failed.
11. To determine teacher competency	*Teachers are competent (or incompetent) but this hasn't been proved. *Performance-based certification is on the way.
12. To help improve teacher training programs	*Teacher training programs need to be more responsive to the needs of the schools.

	*Information to help modify teacher training programs will be readily available.
13. To identify curriculum areas needing improvement	*Curriculum modifications are not keeping pace with the changes required today.
	*The state is considering a state-mandated curriculum.
	*Assessment will cover all major curriculum areas.
14. To form an important element of a state accountability program	*An accountability program could duplicate an assessment program.
	*Accountability has been mandated and all resources must be focused upon it.
	*Assessment is here to stay.
15. To be an important basis for a state PPBS program	*Student achievement indicators have been missing from the current PPBS program.
	*The development of a comprehensive management information system is underway.
	*This is a quarterback sneak to get resources for the PPBS program.
16. To replace the state testing program	*Districts are going to have to pay for their own tests and scoring.
	*Individual test results are only useful as a diagnostic tool in the local school.
	*National tests don't meet our needs.
17. To replace the state accreditation program	*Stricter state standards are soon to be developed.
	*A move is underway to pull out of the regional accrediting association.
	*The state accreditation program is still input oriented.
18. To avoid a national (federal) assessment program	*The current National Educational Assessment Program operated by the Education Commission of the States is a forerunner to a federal program.
	*Rumors about a national curriculum are spreading again.
19. To keep up with the other states	*There may be a failure to examine where the other states are going.
	*State goals may be ill defined.
	*It's time to take another look at the reasons.
20. To gather information so that improvement can be made in education programs for boys and girls	*You're a good man, Charlie Brown!

Implications Based on Operating the Program

With reasons and motives for assessment selected, five general topics associated with operating the program will be explored for implications. These are: (1) control and management, (2) resources – time and money, (3) people, (4) the assessment itself, and (5) reporting.

Control. The implications drawn from the type and level of control of the assessment program will depend to some extent on how the program originated. If the state board of education and the state education agency initiated the program through their own efforts, by administrative action or by proposing enabling legislation, then their leadership role is likely to be secure. If some zealous or *disgruntled* legislators specified the program in legislation, largely without the advice or support of the state education agency, then the likelihood of program success could be diminished. The state education agency in this case may choose barely to meet the literal interpretations of the state assessment law, figuring the pressure won't be so great in the future.

The less specific the enabling legislation, the more likely the state education agency will be able to operate the program successfully. For rules and regulations can be added to from year to year, and modified to fit the circumstances. Permissive enabling legislation rather than purely directive legislation may provide just the kind of state interest which can appeal to school administrators and local boards and bring with it long term success.

Most state boards of education have statutory duties to (1) oversee the public schools, and (2) report annually to the governor and general assembly on the condition of the schools. This kind of authority has been sufficient for several states to develop assessment programs. In such instances, the state board of education, through enactment of clear and positive policies or through adoption of reasonable rules and regulations, can provide an effective base for development of an assessment program. The obvious implication is a strong leadership role for a state board of education.

Can the state board itself escape being assessed – especially on its role in conducting the assessment program? Certainly not, if it expects to stay responsive to citizen needs while at the same time displaying a bit of *brinkmanship* to lead the way to solution of controversial problems.

Cunningham, writing for the Designing Education for the Future Project, offered these comments about state boards:

State boards of education need to be examined rather carefully and their functions assessed. The vitality of state board of education leadership varies, but on the whole has not evidenced the

boldness and farsightedness that seems to be in order. (Implications for Education of Prospective Changes in Society.)

Resources. Two major resources needed for any program – time and money – will be important factors in the success of the program. And these resources must be committed in reasonable quantities for the long-term, for statewide educational assessment is not a one-shot effort. A long-range plan with a realistic allotment of time and money is vital.

So the initial expenditure should be for the development of a plan for assessment. Expenses for this could range from a possible low of $15,000 to as much as $50,000, depending upon the circumstances within each state. Resources spent in this developmental phase may actually save thousands of dollars in operational costs. Four to six months of time should probably be allocated to this planning phase.

Budgetary requirements should be projected for at least three years; a five-year projection should be made if at all possible. Annual costs will probably vary in each of the initial years, and some of the high "start-up" expenses may even occur in the second or third year.

Average annual costs *at the state level only* will vary greatly. The following chart serves to illustrate the range of possibilities (following initial planning expenditures):

	Small state	Medium sized state	Large state
SEA personnel (permanent or contract)	$40,000	$ 75,000	$150,000
Developmental costs	15,000	20,000	25,000
Data collection, scoring, processing	10,000	50,000	150,000
Analysis and reporting	5,000	7,500	10,000
Totals	$70,000	$152,500	$335,000

Let's examine sources of the money first. If the program will be started or maintained with grant funds the implication will be that the state isn't really committed to the program as long as *soft money* is the sole support. It is very easy to discontinue a program when outside funds dry up. State funds channeled through the regular budget of the state education agency imply that legislators and budget officers trust the agency to operate the program in the best interests of all. If such funds are made available in separate appropriations or by annual amendment of specific legislation, then the legislature may be indicating that it wishes to retain most of the control.

Does *statewide* imply that a program is state financed, and is not

costing local districts anything? Not so, as one knows. Even state assessment plans which provide some funds to districts for staff time, proctors, etc., probably never provide for the full local expense.

Sometimes state funds for the assessment program can only be obtained by curtailing or eliminating other state education programs. The obvious implication is that assessment is more important than the other programs, or perhaps objectives of the other programs have been achieved.

Time allotments for the statewide educational assessment program, or major elements of it, are most difficult to suggest *a priori*. The complexity of factors affecting time varies from state to state and with the scope of tasks attempted. The relative nature of deadlines could lead to implications that:

1. one state's assessment program, developed carefully and cautiously, with adequate resources and maximum involvement by the education community, would become fully operational in four years, or
2. another state's assessment program, developed less carefully and in haste, with inadequate resources and minimum involvement by the education community, would become fully operational in four years.

Selecting the right deadline will be both a political and professional decision, with political and professional implications. Generally speaking, however, the legislators and the public will believe too much time was taken in developing the program; educators will believe too little time was taken.

People. Now for a look at the people involved in and subject to a statewide educational assessment program. Both authorities and common sense say that wide involvement is necessary. The definition of *wide* will be up to each state, as it will depend on the size of the state, the number and kind of interested groups, or the number and size of the groups which must be induced to become involved.

Certainly the critics of assessment must be invited to express their views in planning sessions. Their concerns will provide valuable tempering to the plan as it is developed. To not involve the critics will leave the implication that all decisions have been made (or will be made) by the insider advocates and state education agency personnel.

Wide involvement is called for too, so that more will be better informed and educated about the statewide educational assessment plan. Allotting travel expenses for persons to be involved in planning and work sessions will be the best indication that their advice is sought and respected.

Numerous choices are open for the kind and number of professional staff members needed, and all choices depend upon the kind

of program to be developed and the time and money available. Excessive use of out-of-state consultants can erode the leadership position of the state education agency, or cause instate higher education staff to feel miffed. There may be no other way, however, if time is short through legislative mandate. Assessment specialists brought in on a consultant basis should not only assist in the planning, but offer any needed training for state and district personnel.

A state law or board policy directing that a state assessment unit be established in the state education agency has advantages and disadvantages. If such specific direction is given, the implication is that assessment is here to stay, and so it must be visible and be institutionalized. Whether or not the specific unit on assessment is established, the state education agency must have some permanent staff expert on this topic who can direct or coordinate the program. (Be prepared for NASEAMSEA, the currently fictitious acronym for the National Association of State Education Agency Managers of Statewide Educational Assessment.)

Use of knowledgeable school district personnel including teachers as paid consultants (where legal) in the planning and operation of the program will provide untold dividends. If it is not legal to pay them, then they can be given temporary state assignments by cooperation with school districts. The implications: district personnel can do the job, and talent will be around to help evaluate and modify the assessment program in five years.

The largest group to be involved, students, should not be overlooked. This group will probably view assessment as *just some more tests.* Perhaps the most appealing reason for some to participate willingly is the knowledge that assessment is not a *pass-fail* task, and that the results will not be used to rank individuals. The older students' awareness of the purpose of the testing will probably leave most unimpressed, for if there are to be changes made in the curriculum or in teaching methods because of the new *decision-making* information, the changes will come too late to serve the student who is being assessed.

This particular point deserves further consideration. Will the assessment reveal the student's performance in terms of today's or tomorrow's objectives, or yesterday's? Certainly there are few who would want to measure students' abilities to cope with the world of 20 years ago. But is it possible to devise measures which will portend the students' potential 20 years from now? Many talk about educating youth for the world of tomorrow; perhaps it is enough to educate them for the world of today. The implication is that the assessment measures of today are just that, and that to assess for tomorrow

isn't possible. Who knows the future well enough?

Certainly a partial answer to this is that the assessment measures will be modified, changed, and improved each year. If not, another implication arises – that the whole assessment program would tend to set the standards for student performance, curriculum offerings, in fact the whole education program. The creation of such absolute standards could certainly lead to social stratification alien to the traditional purposes of education. With the possibility of fairly rigid standards through institutionalization of the program, the school districts might routinely prepare students for the assessment. And teachers could be trapped into teaching for the test.

The Assessment. Now to the assessment itself. Assuming the question of *Why assess?* has been answered, we find the unanswered questions to be grouped under four *W*'s and an *H*. The major questions: *(1) what is to be assessed, (2) who will be assessed, (3) when will the assessment occur, (4) where will it occur and (5) how will it all be operated.*

Under *What is to be assessed* choices such as these arise: What is the language proficiency of 10-year-olds in our state? Do our seniors possess sufficient arithmetic skills to enable them to make change and complete a sales slip? What are the problem solving abilities of eighth graders? What are the general feelings of sixth graders with regard to persons of ethnic or racial backgrounds different from theirs. What are the average physical skills of tenth graders?

The choices of what are to be assessed lead to a major implication: If the results indicate deficiencies then action should be taken to erase the deficiencies. That's a mighty big order. For since the assessment is conducted from the state level, and since education is a state function, the state, most notably the state board of education, must be prepared to respond. There is always the implication too, that someone had a pet categorical program in readiness, and will be using the assessment to obtain information to justify legislative enactment of the program.

Who will be assessed, where will they be assessed, and when will they be assessed are all dependent upon the reason for assessment and resources available. If the assessment program is going to examine persons not in public schools, e.g., dropouts, graduates, private school students, institutionalized youth, then the implication is that those in charge really want a statewide assessment, one which will provide useful information to other groups and agencies.

The size of the state sample will also determine if every school district will be involved, and to what degree. If the scope of the program permits the full range of instruments to be administered in

every school district, then the implication is that district comparisons are fairly important, and that they will be made. Some districts will certainly seek to have the full range of assessment (even if additional testing must be done) while others will be content with being only in the regular sample, so no district comparisons can be made.

How the assessment will be operated, and what instruments will be used raise a host of implications. Construction and validation of items tailored to a state's objectives for education implies that the state's objectives are somewhat unique and that it is important to see if they are being reached. Use of ready-made norm referenced instruments doesn't necessarily indicate the opposite. Indeed, it may demonstrate that getting the assessment underway fast is important. The use of instruments with national norms will of course permit several kinds of comparisons to be made between your state and region and others.

The ideal assessment program might utilize both testing service and state-developed norm and criterion reference instruments to measure skills and attitudes in the *cognitive* and *affective* domains.

Reporting. Reporting and using the results of a statewide educational assessment program must be carefully planned. The results will be open to hundreds of interpretations, and excessive use of numbers tends to make the results less meaningful to many. The reports must be clear and concise, for this is the decision making information so eagerly sought. A quick look at the kinds of information and the major users will reveal several implications.

Kind of Information	*Major User or Audience*
1. Student	Teachers, parents
2. School and district	Local boards, school administrators, and state education agency
3. Regional and state	State education agency, public, legislators, other states

The implication is that the assessment information would be available and useful to all the groups listed in the right hand column. But this is not a possibility, unless every student is included in the assessment. Care in identifying the major user or audiences is essential, according to Womer, who states:

> *It is important to gather information that will be maximally useful to someone, the legislator, the CSSO, superintendent, the teachers, the pupils, the parents. It is not possible to provide maximally useful information to each of these audiences . . .*
> (Womer, Frank B. *Developing a Large Scale Assessment Program.)*

The nature of the assessment report released at the state level will have a definite effect on subsequent actions. Consider the following sharply worded status report on education in a state:

It is obvious that neglectful school-committees, incompetent teachers, and an indifferent public, may go on degrading each other until the noble system of free schools shall be abandoned by a people, so self-abased as to be unconscious of their abasement. (Nolte, M. Chester *An Introduction to School Administration: Selected Readings.*)

The year was 1837 and the author of the statement above was Horace Mann, secretary to the Massachusetts State Board of Education. This is cited here also as a reminder that educational assessment is not a new endeavor. Perhaps if the style and vigor of Mann's annual reports had been followed by all states through the years, today's demand for assessment programs would be unnecessary.

Summary

In this chapter some choices have been presented to state boards of education regarding the initiation of statewide educational assessment programs. Reasons for assessment programs and some operational aspects of assessment programs were explored and were considered. Efforts were made throughout to cite implications of the choices and actions which may be made or taken by the state board of education and others responsible for the quality of education in a state. *That's what it's all about* — the betterment of public education for boys and girls. Whether this happens may depend upon the decisions made by state boards of education about assessment programs.

CHAPTER III

LEARNING

What Is Learning?

Man learns. This skill, coupled with the means of communication from one human to another, are the reasons why men, rather than some other species, rule the world. Many animals are larger, stronger, and more agile than man, but man, through the power of his intellect and his ability to communicate with other mortals, prevails.

The above supporting illustration surely comes as no surprise to the reader. In fact the reader most probably has regarded this as a self-evident truth for a long time. Why mention it here? Answer, because it's basic. Learning, especially human learning, has been a prime psychological area of rigorous investigation since the science of psychology was founded less than a century ago.

Psychology as a science represented a merger of bodies of knowledge from both the natural sciences (physics and chemistry) and philosophy. With this as a background it was not too surprising to find some psychological investigators (notably E.L. Thorndike) conducting studies into human learning in an attempt to derive some basic laws for learning.

These include:

Learning Law	*Definition*
1. Primacy	Thing learned first is best remembered.
2. Frequency	Thing learned most often is best remembered.
3. Recency	Thing learned last is best remembered.
4. Effect	Man learns what is pleasurable or satisfying to learn.

Lincoln's Gettysburg Address provides the reader with some evidence that Laws 1 and 3 could have some relevance. For example, if the reader is over 30 years of age, chances are that in his secondary school literature experiences, he or she was required to commit Lincoln's Gettysburg Address to memory. You were probably also required to store this *piece of literature* in your memory so it could be instantly retrieved for recital to your literature teacher. Let's see how well the law of primacy applies. If provided with the cue *four score and seven years ago*, you can most probably respond *our fathers brought forth upon this continent a new nation conceived in liberty and dedicated to the proposition that all men are created equal.* If you could make this response, then Thorndike would say that the law of primacy was operating (things learned first are best remembered). Now, if you were provided with the cue . . *that we here highly resolved that these dead shall not have died in vain; that the nation shall, under God, have*

a new birth of freedom, and . . . you will also respond *that government of the people, by the people, for the people, shall not perish from the earth.* Similarly, if you could provide the famous words which conclude the Gettysburg Address, Thorndike would postulate that the learning law of recency had been operating (things learned last are best remembered).

The law of frequency, unfortunately, has been learned too well by the people who prepare advertising scripts for television commercials. Multiple reinforcement, which the law of frequency postulates, requires that the television viewer be exposed to a series of pictures of men with the *wet look* and the *dry look*; more than one housewife who has had her Clorox taken away from her; as well as the two freaky plastic men who have differing amounts of aspirin products in the brain and stomach. When these multiple reinforcement commercials are repeated, there is a great tendency to use this commercial television time to go to the refrigerator for a snack, empty ashtrays, etc. The television advertiser really doesn't care, because once you have seen again the first few seconds of his thirty-second commercial the rest of the commercial and especially the product have been brought to active consciousness, and you will be more prone to purchase that product when next you are shopping.

Thorndike's law of effect states that man has a greater tendency to learn when the thing to be learned is perceived as being pleasurable or satisfying. The law of effect tends to view man as being hedonistic, that is, pleasure seeking, pain avoiding. The extent to which people devote their time and talents to hobbies tends to support this *law*. In-school learning is helped when the room temperature, ventilation, color of the classroom walls, chalkboard, etc., are comfortable and pleasing. Yet these same effect qualities in and of themselves do not cause learning to take place.

Thorndike's laws of learning have not withstood the test of time and further psychological inquiry. Modern learning psychologists discredit the first three laws (primacy, frequency, and recency), and only ascribe partial credence to the fourth (effect). The truth of the matter is that psychologists do not know what learning is!

What people know before a particular learning sequence can be assessed either verbally, through pencil and paper tests, or by asking the person to perform some psycho-motor task. Environmental conditions predicted to be helpful for learning can be provided. Information can be presented on a logical, experiential, developmental or other basis so the learner progresses from the *known to the unknown.* After learning has occurred, the student again can be tested to determine the extent of his learning, but what happens within a person's body (preferably in the gray matter in the human brain) at the instant

or duration of learning remains a psychological mystery.

Psychologists have defined learning as *a change in behavior modified by experience*. Note that the psychological definition is more in terms of what learning does rather than what learning is. Since the exact nature of learning remains unknown, most psychologists tend to regard it as a process. In the most recent psychological literature, learning is treated as a theory rather than with laws or principles. In science theories are neither true nor false but gain value in terms of their usefulness and predictiveness.

At this point a state school board member may be somewhat disconcerted to discover that public education, dedicated to the business of learning, is not only founded on an unprecise science but also on a somewhat uncertain art. School professionals have grown comfortable operating with this uncertainty. They merely assume validity for their theories and procedures, often pointing to traditions as verificity. Public acceptance of these traditions aids in their continuity, although this acceptance is now being challenged. Regardless of how state board members view their present roles with respect to student learning in their states, two facts emerge: (1) at this time no one or group of humans knows many of the answers of how students learn; and (2) state board members have the opportunity to influence student learning in the states, if and only if, they are informed of the most up-to-date trends in educational studies of learning and are willing to be agents for change in the public schools.

Kinds of Learning

In this section the kinds of learning will be examined together with educational applications of each. Essentially three kinds of student learning are of interest to educators. These are:

Kind of Learning	*Definition*
1. Cognitive	Learning which *deals with the recall or recognition of knowledge and the development of intellectual abilities and skills.*
2. Affective	Learning which *describes changes in interest, attitudes, and values, and the development of appreciations and adequate adjustment.*
3. Psycho-motor	Learning which deals with the *manipulative or motor-skill area.*

At the risk of oversimplification it might be said that more than one kind of learning is involved in the student acquisition of any school (or subject matter) related concept. For example, the cognitive component (kind) of learning would be involved in having each student understand the number facts of multiplication of single digit numbers one through nine. Influencing this cognitive component

would be the student's attitudes, appreciations, or interest in this subject matter. Manipulative skills, such as those taught in vocational or home economics courses, along with those skills acquired in gymnastics or other physical education courses certainly would involve psycho-motor components in addition to either a cognitive or affective component or both.

In the more traditional approaches in vogue during the first half of the twentieth century, learning in public education appeared to operate on the sponge theory. In the sponge theory of learning, the teacher, as a master learner, was regarded as a very pliable sponge filled with the fluid of learning. Students would have their sponges moistened by the liquid of learning from the teacher. From a 1970 perspective the description of the sponge theory sounds somewhat obtuse – and it was. Under the sponge theory, learners were relegated to a subservient position in the teacher-learning process. The student was not encouraged to become an independent, inquisitive learner able to pursue learning which may not be within the confines of the teacher's core of experiences.

Other early procedures assumed the mind was like a muscle; if exercised it would grow smarter in capacity. Another theory, one still held widely by the uninformed, assumed the mind had abstract capacities such as logic and reasoning which could be developed independently and transferred to other disciplines. All three of these theories dominated curriculum construction and brought about a tedious, repetitious, elitist, and autocratic school program.

More modern learning methodologies are based upon a series of educational assumptions:

1. Each learner is a unique human being possessing some of the common competencies of his peers in addition to his personal talents.
2. Each learner should be afforded opportunities to develop educationally to the limits of his or her varied abilities.
3. Each student should be afforded a variety of approaches for learning which best matches the student's learning style.

The learning style referred to in the third assumption is related to the specific approach which each learner finds most advantageous to his learning. The learning style not only varies from child to child, but also from one subject matter to another for the same child. Some learning styles are visual only (reading and seeing), others auditory, while still others require some combination for the most effective learning.

Cognitive Learning. The content of most subject matter areas taught in public schools can most appropriately be referred to as cognitive learning. Levels of cognitive learning include all the knowledge, un-

derstanding, application, analysis, synthesis and evaluation of the facts, principles or concepts to be learned in a given unit, quarter, or semester in a given course. According to the *Taxonomy of Educational Objectives, Cognitive Domain*, these levels are defined as follows:

Level of Cognitive Learning	*Definition*
1. Knowledge	*the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting.*
2. Comprehension	*understanding or apprehension such that the individual knows what is being communicated and can make use of the material being communicated.* (Ability to visualize the relationship between and among concepts.)
3. Application	*the use of abstracts in particular and concrete situations. The abstractions may be in the form of general ideas, rules or procedures, or generalized methods. The abstractions may also be technical principles, ideas, and theories which must be remembered and applied* (problem solving).
4. Analysis	*The breakdown of a communication into its constituent elements or parts such that the relative hierarchy of ideas is made clear and/or the relations between the ideas expressed are made explicit* (deduction).
5. Synthesis	*The putting together of elements and parts so as to form a whole. This involves the process of working with pieces, parts, elements, etc., and arranging them and combining them in such a way as to constitute a pattern or structure not clearly there before* (induction).
6. Evaluation	*Judgments about the value of material and methods for given purpose. Quantitative and qualitative judgments about the extent to which materials and methods satisfy cri-*

teria. Use of a standard of appraisal. The criteria may be those determined by the students or those which are given to them.

Those levels of cognitive learning are usually contained in the educational objectives developed by the teacher for a given lesson, for a unit, or for a course. Depending upon the complexity of the concepts, or procedures to be acquired by the student, the teacher will vary the methodology of presenting the information to be learned. For low level knowledge the teacher perhaps will present the information verbally to a group in class. The teacher may choose to employ examples, comparisions, statistics or testimony of authorities in support of the main ideas. This support material could also take the form of a film strip, motion picture or audio tape in addition to the teacher vocalization. If the teacher were attempting to individualize the student learning in her class, she might choose to present the class with a series of learning alternatives to acquire the concepts. Using this methodology, the teacher is the manager of the learning situation. The teacher works closely with each learner either singly or in small groups when more than one learner selects the same learning approach and chooses to work with other students.

Teacher methodologies tend to change from lesson to lesson, course to course, and from the early elementary to the higher secondary grades depending upon: the complexity of the concepts to be learned; the sophistication of the student; and the available support (audio visual) materials which the school district can bring to bear to assist the teacher. Most learning content in grades K through 12 represents cognitive learning.

Affective Learning. The content of learning carries with it, aside from the mental gymnastics of the cognitive domain, an affective component in the mind of the learner. Each learner tends to like, to dislike, or to be ambivalent. to all that he learns. Additionally, society demands that succeeding generations acquire positive attitudes toward law and order, democracy, and moral and ethical values. Levels of affective learning include: receiving; responding; valuing; organization of a value; and, characterization by a value. In the *Taxonomy of Educational Objectives: Affective Domain,* these levels are described as follows:

Affective Learning Level	*Description*
1. Receiving	*The learner be sensitized to the existence of certain phenomena and stimuli; that is, that he be willing to receive or attend to them.*

2. Responding	*This is the category that many teachers will best describe their interest objectives . . . . indicates the desire that a child become sufficiently involved in or committed to a subject, phenomenon, or activity that he will seek it out and gain satisfaction from working with it or engaging in it.*
3. Valuing	*. . . that a thing, phenomenon, or behavior has worth. This abstract concept of worth is in part a result of the individual's own valuing or assessment, but it is much more a social product that has been slowly internalized or accepted and has come to be used by the student as his own criterion of worth.*
4. Organization (of a value)	*As the learner successively internalizes values, he encounters situations for which more than one value is relevant. Thus necessity arises for (a) the organization of the values into a system, (b) the determination of the interrelationships among them, and (c) the establishment of the dominant and pervasive ones.*
5. Characterization by a value	*. . . the values already have a place in the individual's value hierarchy, are organized into some kind of internally consistent system, have controlled the behavior of the individual for a sufficient time that he has adapted to behaving this way; and an evocation of the behavior no longer arouses emotion or affect when the individual is threatened or challenged.*

Teachers are usually involved with affective learning up to and including Taxonomy level three (valuing). Professional testing companies are capable of developing tests to include levels four and five. The affective component of the cognitive learning content is of great interest to teachers, since students tend to learn better those things which they like or have positive attitudes toward. Courses in civics, group discussions on *what would you do if*, situations in public affairs topics or the student's comments relative to the practical ap-

plication of vocational or home economics topics to real life situations similarly need to be known and capitalized upon by the teacher. The attitudes a student leaves school with are the result of many years of attitude and value formulation. Teachers and school systems can seek to foster positive attitudes on the part of their students if the teachers or other school officials are willing to discover student feelings and attitudes at all points along the way in the educative process. School personnel must also be willing to develop corrective strategies and approaches to overcome those educational practices which unnecessarily provoke unfavorable student feelings, emotions and values.

Psycho-Motor Learning. Psycho-motor learning involves those skills, procedures or operations which involve some manipulation of objects or the body itself. These learning activities are most often associated with handwriting, speech, physical education, industrial arts, and technical courses. Although the manipulative component of psycho-motor skills is assessed by some sort of performance indicators, most teachers tend to regard the course content as they regard other cognitive student learnings. Thus, while a demonstration-performance methodology (wherein the teacher explains and demonstrates the skill, the student performs under teacher supervision, and the student performance is evaluated as to accuracy of procedure and outcome) is fairly common to psycho-motor aspects of physical or vocational education courses, teachers continue to depend primarily on cognitive processes to ascertain overall student learning.

As teachers, curriculum specialists and school administrators isolate and group concepts to be learned into lessons, units and courses of study, objectives identifying these competencies are written for use by the classroom teacher. The specific methodology each teacher will employ to assist students in learning these concepts will be partially dictated by the nature of the content of each educational objective. The teaching of methodology also is influenced by the specific nature of the learning environment. Self-contained classrooms, team teaching, individualized instruction, departmental organization (usually at the secondary level), modular scheduling, etc., each have methodological advantages and limitations for the teacher.

The remaining sections of this monograph will be devoted to: the measurement and evaluation of student learning; the problem of *appropriate educational criteria*; some suggestions for reporting the outcomes of evaluating student learning; and some rules of thumb which state school board members might employ to assist them in the evaluation of student learning resulting from curricula under their purview.

CHAPTER IV

MEASUREMENT

Measurement is the art of quantifying something. E.L. Thorndike in the pioneer days of the testing movement the United States was known to have made the very modest comment, "If something exists, it exists in a quantity. If it exists in a quantity, I can measure it." Measurement may take one of several forms. When a person visits his physician he is weighed, has his height taken, has his blood pressure taken. He has his urine and blood analyzed for specific gravity and composition. The doctor then records these measurements, and by comparing these recordings against some standards, judges (evaluates) the condition of the patient's health.

The same is true in educational measurement. The purpose of educational measurement is to assess whether the student has attained (learned) the various objectives of the curricula (for practical purposes in education the words "measurement" and "assessment" are synonymous.)

What kinds of measurement do educators employ in their trade? Certainly many are used. The nature and form of the measures vary according to the requirements of educational objectives. The physical education teacher may employ a modified check list which he annotates every time a student masters a given feat in gymnastics. Another physical education teacher may record the number of pushups, pullups, situps, etc. and measure the progress each student makes throughout the course.

The social studies teacher may well employ a sociogram to measure the group dynamics of her class. Through this technique, the stars (very popular students), cliques, and the isolates (students who don't have friends in the class) are determined. The social studies teacher also may employ an affective questionnaire to measure the class feelings toward race, government, and social issues.

The English teacher tends to employ student essays to measure both the content which students have learned and also the students' ability to organize their thoughts coherently in written communications. The English teacher also uses the verbal questions and answers technique with students in the classroom to measure whether the students appear to be learning.

Problem solving formats are employed by virtually every teacher to ascertain whether students have mastered a given set of operations. This is true in the demonstration of mathematics procedures as it is also true in logical presentation of points of view in a *describe and compare* type of essay question in Civics.

Virtually all schools employ objective tests (either nationally or locally prepared) to measure the student learning in such areas as reading speed and comprehension, math concepts and computation, English usage, History, Social Studies, etc.

These various educational measures are recorded by the teacher on either an absolute scale (percentage or letter-grade) or on a relative basis (a student's position with respect to some national or regional norm as the 60th percentile).

Kinds of Educational Measurement

Usually four kinds of educational measures are found in public school programs, (1) intelligence tests, (2) achievement tests, (3) aptitude tests and (4) diagnostic tests.

Intelligence Tests: These instruments purport to measure the student's intellectual capacity. The two major kinds of intelligence tests are (1) group and (2) individual. As the labels signify, a group intelligence test can be administered to a class under the guidance of a teacher. An individual intelligence test (Stanford-Binet, Wechsler) is administered to a student by a psychologist on a one-to-one basis. Scores resulting from these tests vary. Some tests yield a total score only, while others yield scores for verbal, quantitative and total. The obtained score then is converted into an intelligence quotient. L.M. Terman of Stanford University in 1916 first described the Intelligence Quotient as the ratio between mental age and chronological age ($IQ = \frac{\text{mental age}}{\text{chronological age}} \times 100$). Psychological studies of intelligence tests tend to reveal that they are culturally and ethnically biased in favor of the white, middle class American. Intelligence tests are often used to predict a rate of expected student growth. The child whose IQ is 100 usually progresses 10 months in grade achievement (one school year) per year, the 80 IQ youngster typically will progress 8 months in grade achievement, etc.

Achievement Tests: Achievement tests are designed to measure a student's educational progress in comparison with other similar students nationwide. Such measures as *California Achievement Test, Iowa Test of Basic Skills, Metropolitan Achievement Test, Sequential Tests of Educational Progress, Stanford Achievement Test, Test of Academic Progress, Wide Range Achievement Test*, and others, contain subtests which assess reading, English Usage and composition, Math, History, Social Studies, etc. The subtest scores are usually converted into grade equivalent scores. A student with a grade equivalent of 5.1 is said to be functioning at the first month (October) of grade 5. (The decimal portion of the scale divides the 10-month school year such

that September is 0, October is 1, November is 2, etc., until June is 9). The basis for these grade equivalent scores stems from an analysis of a national sample of students. Tests such as these will undoubtedly backbone any statewide assessments.

Aptitude Tests: Aptitude tests are designed to measure a student's propensity for a given course of study (say geometry) or an occupational field of work. The tests are designed such that a person scoring high on the test would also be likely to be successful either in a given course or field of work. Aptitude tests are prerequisites for entrance into most medical, dental, law, engineering, and teaching schools. Tests such as the *Miller Analogies Test* and the *Graduate Record Examination* are required for graduate study. The *Strong Vocational Interest Blank* and *Kuder Preference Record* provide valuable information which students use (with educational guidance) to select the next steps in their career planning.

Diagnostic Tests: Diagnostic tests are employed to discover the nature and degree of student learning difficulty. Standardized tests for reading and math diagnosis are available for the normal student who experiences difficulty in one or both of these areas. For the special education youngster whose IQ is less than 90, a whole battery of standardized instruments is designed to diagnose the specific learning dysfunction. Among these are:

Bender - Gestalt;
Frostig Test of Visual Perception;
Wepman Test of Auditory Discrimination; and
Illinois Test of Psycholinguistic Abilities.

It should be noted each kind of measuring instrument just described was designed for certain educational purposes. The procedures to validate each of these measures to gather large groups of representative students, and to derive the interpretation of the test results, are both rigorous and exacting. It is toward an examination of the characteristics of a good measuring device that we now turn our attention.

Characteristics of a Good Measuring Device

An effective measuring device should have the following five characteristics: (1) validity, (2) reliability, (3) objectivity, (4) comprehensiveness, and (5) differentiation.

Validity: A test is valid if it measures what it was intended to measure and nothing else. The American Psychological Association's *Technical Recommendation for Psychological Tests and Diagnostic Techniques* (1954) identifies four categories of validity. These are content, predictive, concurrent, and construct validity.

Content Validity: *Content validity involves essentially the*

systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured (Anastasi, 1966). At first glance the determination of content validity would seem to be a rather easy task. The *hooker* comes when one has to say that the questions cover all major aspects of the course content in the appropriate proportion. Some areas of learning lend themselves quite readily to objective questions, e.g., true-false, multiple choice, matching. It is considerably difficult to generate questions for concepts which are abstract and global. Another concern is, does the test (usually achievement) measure the objectives of the instructor as well as the objectives for subject matter? To check statistically for content validity it may be necessary to prepare parallel forms of the test. Using this procedure one form is given as a pretest, and the second as the posttest. The gain made in score provides some evidence of its content validity.

Predictive Validity: *Predictive validity indicates the effectiveness of a test in predicting some future outcome* (Anastasi, 1966). A student's scores are checked against some measure of success on a given job or course of study. The *on the job* performance is usually regarded as the criterion. Predictive validity of a test is required before the instrument can be used as a screening device for hiring job applicants, selecting students for admission to college or professional schools, etc.

Construct Validity: *The construct validity of a test is the extent to which the test may be said to measure a* theoretical construct *or trait. Examples of such constructs are intelligence, mechanical comprehension, verbal fluency, speed of walking, neuroticism, and anxiety* (Anastasi, 1966). Construct validity of a test can be ascertained over time. For example, with respect to intelligence testing, did the ratio of mental age to chronological age remain stable as the student got older? Other statistical techniques employed to assess the construct validity of a test are: (1) the correlation of this test with scores on already validated tests which measure the same traits; and. (2) factor analysis, which groups the test items according to traits or factors being measured.

Reliability: *The reliability of a test refers to the consistency of scores obtained by the same individuals on different occasions or with different sets of equivalent items* (Anastasi, 1966). The two major types of reliability estimates are determined by either test-retest correlation or by an internal consistency index. In test-retest reliability (coefficient of stability), the test is readministered to the same students after a time interval and a correlational coefficient is computed. In computing an internal consistency reliability in-

dex (coefficient of equivalence) the test is divided in half, by making a subtest of odd-numbered items and another subtest of even-numbered items or by making subtests of the first and second halves of the test, and correlation for one half of the test, is computed.

Objectivity: A test has objectivity if the same score is assigned no matter who corrects the test. Thus, true-false, multiple-choice, and matching items lend themselves to objectivity; essay items do not.

Comprehensiveness: A test is comprehensive if the test items adequately sample the full range of objectives and subject matter in a curriculum. Careful attention needs to be paid, lest trivial information items account for more than their proportionate share of the test. So while the test is designed paying careful attention that all aspects of the educational objectives and course content are included in an instrument, it must be assured that the number of items pertaining to a given content also be consistent with that content's importance.

Differentiation: Items on a measuring instrument have differentiation when the brighter students get the item correct, and the less intellectually talented students get the item wrong. The same thing can be said for the test as a whole. In item analysis, the ease index is the ratio of the number of students who scored correctly on the item and the total number of students who took the test

(E.I. = # of students correct/total # of students)

Professional test writers seek items having ease indices in the .40 - .60 range for achievement tests. It should be noted that differentiation may not always be desired, especially for locally prepared achievement tests. Some educational objectives require mastery level learning by all students. The major concept in a course, for example, might require student mastery before extensions of these concepts can be learned. When local achievement tests are designed, items measuring the student learning of *mastery material* should have ease indices .85 and higher. So, a note of caution is suggested when critiquing a test for discrimination. The critiquer must ask *Does the objective that this question is written against require mastery learning by the student?* If the answer is *Yes* then an E.I. of .85 or more is to be expected. If the answer is *No* or *not necessarily* then E.I.'s is the .40 - .70 range are to be anticipated.

It should be noted that the characteristics of good measuring devices apply not only to commercially developed instruments, but also to those generated locally. The educator's purpose for measurement, as well as corresponding interpretations, are the

topics to which we now turn and will complete our examination of educational measurement.

Norm-Referenced vs. Criterion-Referenced Measurement: The standardized measuring instruments described thus far in our section on educational measurement are best described as norm-referenced instruments. The norm-referenced label refers to the purpose of developing the instrument in the first place, and the interpretations which can be made after a student has taken the test. Take, for example, the Stanford Achievement Test. The Stanford Achievement Test has been developed and validated to reflect what the test developers felt were common elements of national student learning for reading, English, science, social studies and math for respective elementary grade levels. The various sub-scores attained by a student are transformed into *percentiles* or *grade equivalent* scores so a given student's progress can be compared to the attainment of many other similar students nationally. This practice has the advantages of: (1) being able to see how individual students rank on a *national standard*; (2) getting a feeling of the relationship of the local curricula to that which is being taught nationally to similar students; and, (3) providing some accountability information to the local public which is subsidizing public education. Norm-referenced instruments may or may not be relevant to either the number of objectives or the relative emphasis of the various educational programs.

How then does an educational administrator assure that the testing program measures in fact the relevant educational objectives of his programs? At least two alternative solutions appear possible: (1) the local school district could adopt the objectives of the norm-referenced instrument, and go with a national curriculum - which is not too tenable; or (2) they could develop their own local achievement test which measured their objectives. To do the latter, the administrator must resort to some form of a criterion-referenced testing program. A criterion-referenced instrument is one whose questions are specifically tailored to measure the educational objectives of the local school program, and a level of attainment is specified. The criterion-referenced testing movement has been gaining momentum in American education since the mid-1960s. In criterion-referenced testing it is not only possible for a student to achieve a perfect score, it is desirable. Criterion-referenced testing (where the local goals and objectives are the criteria) provides an ideal solution to measuring student learning which is relevant to the local educational program, but this solution is not without its problems.

First, the construction of measuring instruments is a highly specialized skill which few local educators have had either academic training

to develop or the experience and expertise to analyze and norm the results.

Second, some of the national apostles, disciples, and followers of the criterion-referenced movement reject the necessity for reliability and differentiation as necessary prerequisites for criterion-referenced instruments under the guise that these instruments should reflect mastery learning. This second *problem* will be elaborated upon in the criteria section for reliability, but for now, it suffices to say that making measuring instruments relevant to local curricula, in no way means that one should discard the time-honored and validated requisites for good test construction. The notion that criterion-referenced measuring instruments should measure *mastery-level* objectives is similarly absurd. The American public educational program is based upon the notion that each child has the opportunity to learn to the extent of his or her talents. Certainly the notion of individual student differences would dictate that the criterion-referenced measuring instrument would provide challenge for the gifted child, or the child whose talent is centered in a given idiosyncracy (say, math, English, social studies, etc,) also would have his learning measured.

Third, the local school district, even after it had developed and validated its criterion-referenced instruments, still requires some data relative to the comparability of its findings, with other school districts in the state or the region. This problem makes a strong case for having district-to-district communication operating within a state or a geographic region of the country. Each student completing a grade K-12 public educational program matriculates into the next step in his preparation for a career, Whether the next step for the student is college, community college, technical training, or the world of work, grades K-12 learning experiences should have been equally relevant for such a transition.

Thus, while criterion-referenced measurement has much to recommend it, a matrix of educational problems must be overcome to assure that it is being employed effectively. The need for local districts to develop this expertise in criterion-referenced measurement was reinforced in the present session of Congress when a bill requiring criterion-referenced testing was introduced (HR-69) to justify a district's entitlement to Title I, ESEA funds beginning with FY-75. If this requirement is in the final law, considerable progress will be made in criterion-referenced measurement for school districts vying for future Title I funds.

Needs Assessment: Twenty-three of the fifty states have educational accountability legislation. Most statutes require that local school districts design their accountability programs from data gathered by districtwide educational needs assessment. A needs assessment

instrument is one which is designed to measure the school district's educational needs as perceived by educators, students, parents, commerce and business, the trades, and members of the public-at-large. The instrument itself is typically designed to measure objectively the content of the various educational programs in existence in the district. Respondents also have the opportunity to communicate their desires for the direction the educational program should take as well as their critique of the existing program. Local educators then synthesize these data as new goals and objectives for the district are redesigned. Within the recent past, needs assessment has provided educators with a mode of communication with the public which has heretofore not been available for the modification of local educational curricula. In Colorado, for example, virtually all of the 181 school districts conducted needs assessment as an integral part of their local accountability program, and have adjusted the planning of educational programs accordingly. These needs assessment measuring devices have provided local educators with an opportunity to optimize their educational offerings to coincide with the manifest desires of the public which they serve.

National Assessment: In the mid-1960s, the Education Commission of the States, a non-profit organization representing over 40 states and territories, established a subsidiary, the National Assessment of Educational Progress (NAEP). NAEP was designed to measure the educational attainment of a sample of the population ages 9, 13, 17 and 26-35 across a range of educational competencies in an effort to assist local educators in appraising the effectiveness of educational programs on both a national and a regional basis. Assessment instruments are and will be designed to measure the following:

Cycle I

1969-70 Citizenship, Science, Writing
1970-71 Reading, Literature
1971-72 Social Studies, Music
1972-73 Math, Science
1973-74 Writing, Career and Occupational Development
1974-75 Art, Citizenship

Cycle II

1975-76 Reading, Literature
1976-77 Music, Social Studies
1977-78 Math, Science
1978-79 Writing, Career and Occupational Development
1980-81 Reading, Literature

(*Compact*, Feb. 1972, Vol. 6, No. 1, p. 17)

The first few reports generated by NAEP have resulted in mixed reactions by local educators. NAEP is solely an assessment (measurement) activity. The purpose was to assess (measure) the knowledge, understanding, abilities, and feelings of learners toward the subject matters in the schedule described above.

Some local educators have reacted negatively to NAEP reports because: (1) National and regional findings only were reported (NAEP never promised more than this); (2) the age categories represented criteria which tended to confuse the usual *grade level* data with which the local educators were more familiar; and,(3) the NAEP reports contained few, if any, analyses resulting from the assessment other than the presentation of data. (In all fairness to NAEP, assessment rather than evaluation [judgment based on assessment] is their *raison d'etre*, and the local educator's critique should be, but isn't, from that frame of reference.) It should also be noted that educational agencies more regional or national in scope (notably the U.S. Office of Education, NAEP's principal benefactor) have found the NAEP results to be extremely valuable.

Local educators are used to having the data measuring students' achievement presented in some sort of perspective. They expect that these data will provide them with a series of specific instances where their educational programs are succeeding, and where they are not. Usually, local school district administrative staff data are not attuned to the educational objective level (by subject matter) documentation reported in NAEP studies. Thus, while it is interesting for the local educational administrator to know that 62% of the students in his region do not know basic tenets of social studies measured on the 1971-72 NAEP instrument, he must still go out and develop his own social studies instrument to see if his students lack these competencies, even though his district might have been subject in the NAEP 1971-72 study. This, most districts are neither staffed nor willing to do.

NAEP's reporting on a chronological age basis of 9, 13 or 17 years old similarly has caused problems in local educational agencies. Since the beginning of the testing movement in the country, students have been grouped on a *grade level* basis. While it is recognized that for a given grade level, students of differing chronological age are present, the learning achievement of these differing chronological age students could be measured because they had been exposed to a comparable level of course content. NAEP, in their reports, categorize students according to chronological age and report their (the students') learning in a given subject matter area by the age criterion. Local educators are entitled to feel that the NAEP findings present a confounding variable to data which they already have, because this is what has occurred. Again, the local district must assess over multiple grade

levels to determine whether the NAEP findings pertain to their students aged 9, 13 and 17.

The third LEA (local educational agency) objection to the NAEP reports relates to the unwillingness of NAEP to analyze their data locally and make recommendations. As a national subsidiary of the Education Commission of the States, NAEP has felt a need to remain apolitical in reporting their data on a regional and a national basis. As time goes on, and NAEP, as established and functioning as a viable, measurable component of American public education, continues to gain the national, regional and state recognition for their expertise in assessing the state of public education, perhaps, analysis and recommendations can appropriately be made on a state and local basis.

NAEP has made a commendable contribution to measurement of student learning in public education. The foregoing dialectic was in no way an effort to impugn what NAEP has attempted to do to assist in assessing public education nationwide. The reporting parameters have been different, but if NAEP wants to remain in the fore, it must either: (1) provide local educators with the background to cope with and interpret NAEP reports locally; or (2) revise their reporting format so they (NAEP) provide local interpretation and suggestions for improvement. Although neither the purpose or the operational procedures of NAEP require this level of detailed analysis, local educators need this level of detail to modify their existing educational programs. At present, an impasse appears to exist between NAEP and local educational agencies.

CHAPTER V

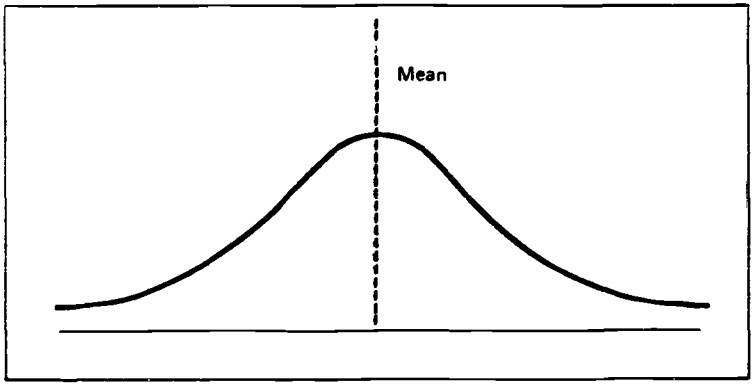
Evaluation

Evaluation is judgment based upon criteria. This classic definition of educational evaluation has withstood the test of over a half-century of use in the field of public education. Educational evaluation is a terminal activity associated with educational measurement. Evaluation is sought to answer the question *So what?* to data. Asking the simple question *So what?* may be a terse verbal overture. The appropriate reply when applied to a diversity of educational applications, is answered in terms of statistical probabilities. Educational measures are not absolute truth but are merely contrasted against the likelihood of chance.

Virtually every human trait (intelligence, height, weight, etc.) when measured is distributed according to the normal probability curve. The normal curve has certain characteristics as shown in Figure 1.

FIGURE 1

The Normal Curve



An inspection of Figure 1 reveals that the curve peaks out around the midway point along the horizontal axis. When the curve is truly *normal*, the point on the horizontal scale vertical to the highest point on the curve describes three statistical measures of central tendency. They are:

Measure of Central Tendency	*Definition*
1. Arithmetic Mean	The sum of all the scores along the curve divided by the number of observations.

2. Median	The midpoint of the curve such that divides the normal curve into two equal areas.
3. Mode	That point along the horizontal axis at which the normal curve reaches its highest point.

The mean, median and mode are employed as the measure of central tendency in most educational statistical computations. The mean is used as a measure of central tendency when the scores are normally distributed and the quantity being measured is being reported on an interval scale. (An interval scale is one which has equal units of measure at all points along the scale, i.e., weight in pounds, height in inches, scores, raw scores on a test, etc., each are examples of interval measurement.)

The median is used as a measure of central tendency when the shape of the distribution is influenced by one or more extreme scores. Under such circumstances the middle score rather than the mean would be the most appropriate measure of central tendency. The following example illustrates the appropriate use of the median rather than the mean in an educational setting.

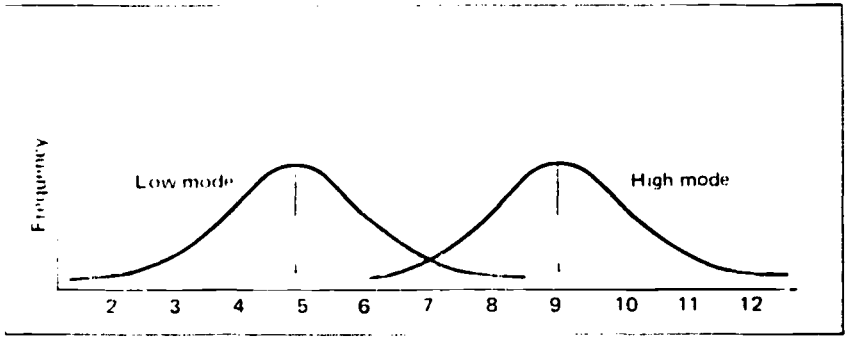
Suppose the local board of education is desirous of knowing the average compensation paid to the professional staff in their small district. The salaries are:

1 Superintendent @	$20,000
1 High School Principal @	$18,000
1 Elementary Principal @	$15,000
7 Teachers @	$ 8,500
6 Teachers @	$ 7,500
3 Teachers @	$ 7,000

If the mean salary ($9,394.75) were reported to the school board rather than the mid-score (median) value of $8,500.00, the reported value would have been inflated by $894.75. Note that the median score is not influenced by extreme values. If the superintendent's salary were increased to $29,500, the mean score would increase $500, while the median score would remain the same.

The mode is the most frequent score that occurs in a given distribution. In a truly normal distribution the mode, median and mean occur at the same point on the scale. Sometimes, more than one modal point occur in a distribution as in Figure 2.

FIGURE 2
A Bi-Modal Distribution

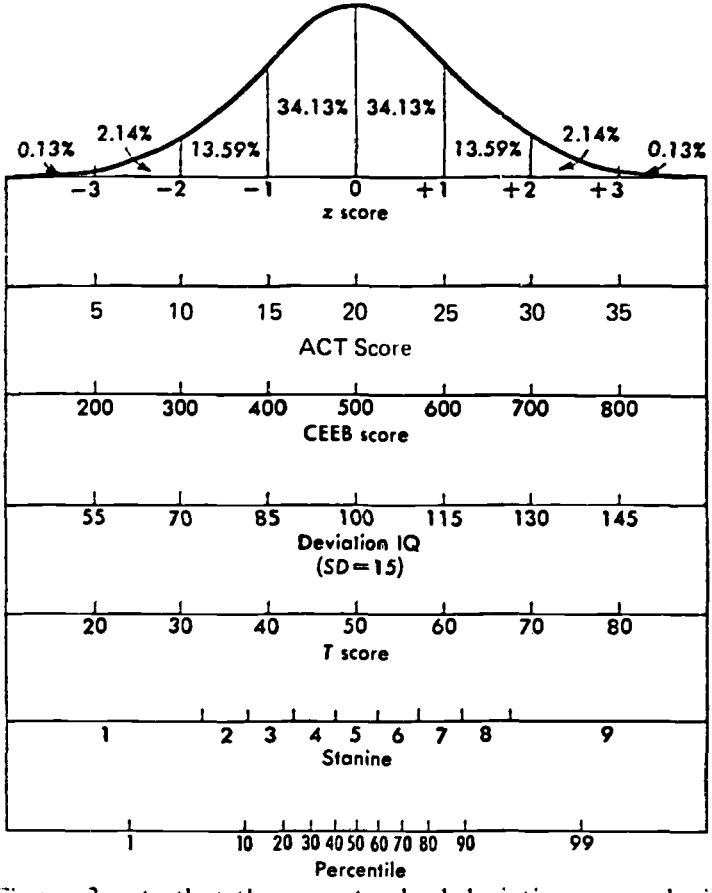


In Figure 2 modal points occur above 5 and 9 on the horizontal scale. As one observes multimodal data, he tends to react that what was measured and displayed was done so on a population having more than one characteristic. For example, if Figure 2 represented data on the numbers of people having different shoe sizes, the bimodal distribution resulted from combining both men and women in the sample displayed. The first modal point encountered (5) most probably is the central tendency for women's foot sizes; whereas, the second modal encounter at 9 represents the central tendency for men's foot sizes.

In summary, there are three statistical measures used to identify the central tendency of a distribution of scores. These are the mean (arithmetic average), the median (middle score) and the mode (the most frequent score.

The normal probability curve also can be divided into a series of equal units along the horizontal axis which partitions the area of the normal curve into a series of standard units called the standard deviation. The standard deviation is a statistical measure such that one standard deviation unit on each side of the mean includes approximately two-thirds of the number of cases in a given distribution. Standard deviation scores are represented by a scale of z scores which has a mean of zero (0) and a standard deviation value of one (1). This relationship is described in Figure 3.

FIGURE 3
Relationships among Different Types of
Test Scores and the Normal Curve



In Figure 3 note that the one standard deviation on each side of the curve includes 68.26% of the area under the curve (or the number of people contained in a distribution of scores). Between +1 and +2 and -1 and -2 another 13.59% is contained respectively. As can be observed the z score partitions the horizontal axis under the normal curve into a convenient number of equal intervals, and provides the statistician with a clear picture of where a given student score places that interval along the scale. For example, if a student's achievement on a test is equivalent to a z score of +2, the evaluator knows that this attainment is two standard deviations more than the mean achievement for all students who took the test. It also places him within the

top 2% of these students. Reporting test results to students and parents in terms of z scores can be somewhat disconcerting. Imagine the chagrin for the parents and the child whose achievement was exactly on the mean. His z score would be zero. Needless to say, the teacher would have to expend considerable time with both the parents and the child to assure that they understood that not only was the child's achievement greater than zero, but that his performance was perfectly acceptable against an average criterion.

While a standard deviation unit scale (often called standard scores) provides valuable statistical information, for reporting purposes other standard scales have been produced. McCall, for example, developed the T score scale which has a mean of 50 and each standard deviation is assigned a value of 10. The American College Testing instrument has a mean of 20 and a standard deviation of 5. The College Entrance Examination Board test battery has a standard score scale which has 500 for a mean and 100 for a standard deviation. The Wechsler Intelligence Scale has a mean of 100 and a standard deviation of 15. It becomes readily apparent that any specialized standard score can be created by arbitrarily assigning values for the mean and the standard deviation.

The stanine scale divides the 100% area under the normal curve into nine divisions. Stanine scores have a range from 1 to 9 with an average of 5. These scores have been valuable in predicting success in pilot training and specific college courses.

The percentile scale represents a cumulative recording of the area under the normal curve. Percentile scores present a relative standing for any individual vis-a-vis his standing in the group of people who also have taken the test. If a student's achievement on a given test places him at the 70%-ile, this means that his achievement was the same or better than 70% of the students on whom the test was normed. From mere observation the reader can ascertain that the percentile scale is tightly packed around the middle and widely extended in terms of numerical intervals on either extreme. Since percentile scores cannot achieve the requirements of interval data, they cannot be added, subtracted, multiplied or divided for use in statistical manipulation. Percentiles, therefore, are widely used in reporting achievement results to parents, because the concept of percentiles is closely related to the notion of percentage, a concept with which parents are familiar.

Norms

When measuring instruments are developed, they are usually field tested on a large sample of students. The sampling procedure em-

ployed by most testing companies is usually both exhaustive and comprehensive. Students are selected for inclusion in either a national or regional sample on either a random or a stratified basis. A random sample is one where every person eligible to be selected has an equal opportunity to be included in the sample. Random sampling represents an underlying mathematical assumption for parametric statistics such as mean comparisons, analysis of variance, etc. A stratified sample is employed if the sampler is interested in obtaining students who have differing desirable characteristics in some proportion. It is quite common to use socio-economic status, sex, ethnic background, etc. as the basis of some proportionate stratification. Within each strata, subjects are randomly selected for inclusion in the sample.

After the students have been selected for inclusion in the sample, and the test has been administered to them, mean and standard deviation values are determined. From the mean and standard deviation statistics, specialized standard scores are developed (e.g., CEEB, ACT, T score and z score) and for each standard score, a percentile equivalent, grade equivalent or similar score is generated for norming purposes. These norms are developed on either a regional or national basis for the various strata employed. Then, when this test is administered in a local school district, each student's achievement can be compared against these regional or national norms.

It should be stated that as an artifact of statistics the use of norms results in decisions such that half the students measured score above and half score below the 50%-ile. In this statistical sense the percentile rank of a given school (or student) makes no comparison relative to the competency level in the subject matter that has been attained.

Criterion-referenced measurement practices tend to report a given school or school district's level of competence against each of the educational objectives which the instrument was designed to measure. If the criterion of success is agreed before testing to be 70%, 75% or 80%, a comparison of the average ease index for the items measuring a given objective provides the evaluator with the data required to ascertain if the achievement met or exceeded the criterion standards.

Ranking students against national, regional or state norms may provide useful information for educational guidance and counseling for that student and, as such, could be a good thing. Realistically, there is little educational benefit to be derived from such normative comparisons for individual schools or school districts. School and school district data should be normed against criteria rather than normative standards. The educational effectiveness of a given school or school district ultimately rests in the educational competencies of its students. This student competency can most validly be assessed against the actual educational objectives in use in the school. Then,

and only then, can some judgment be rendered relative to the educational quality, effectiveness, or the cost of efficiency of a given school district's educational program.

A Point of Compromise

Up to this point, the reader has probably gained the impression that norm-referenced measurement is employed to rate students against some national standing; whereas criterion-referenced testing purports to measure the student mastery of learning objectives contained in a given school district's educational program. This impression is a valid one. A point should be made, however, that the comparison of data yielded from each measurement device need not necessarily result in an *either-or* type of measurement decision. Certainly, a given norm-referenced instrument does not contain items which comprehensively measure the content of a given district's English, math, reading or social studies programs, but chances are that it does measure a significant portion of these respective programs (else why did the faculty recommend the use of a given norm-referenced instrument?). The questions on the norm-referenced instrument could then be related to the school district's program objectives. When this task is completed, some local objectives will have five or more questions, others one or two questions, and still other objectives will have no questions on the norm-referenced instrument.

An analysis of these items and objectives will reveal objectives which have been partially measured and others that have not been measured at all. This analysis then becomes the basis for constructing a criterion-referenced instrument. At testing time, both instruments are administered. The usual norm data are available for use in the traditional manner for each student. Similarly the student answers for both instruments are partitioned by local district educational objective to ascertain whether the obtained ease indices meet the established criteria. This dual approach would enable local educators to conduct their districtwide assessment and relate these data to the effectiveness of their local educational programs. This *compromise* measurement procedure would allow the local district with the lowest possible cost for the criterion-referenced measurement and evaluation of their educational program.

Many school districts who employ norm-referenced measurement also purport to have programs which individualize instruction. Certain *individualized programs* such as Individually Prescribed Instruction (IPI); Individually Guided Education (IGE); Project PLAN (Westinghouse Learning Corp.); etc., each have the notion of student learning as their goals. Norm-referenced instruments can be employed to determine whether a given student is working up to his *potential* through comparing the actual achievement with some

measure of his expected achievement.

In the early 1960s the California Testing Company (now a division of McGraw-Hill) incorporated an intellectual status index into the California Achievement Test Battery. The intellectual status index employed the mental age portion of the student's IQ to project how much progress a given child should make in a school year. The technical conversion grade equivalent measures for each IQ score are then made. The practice then was to compare actual achievement with *expected achievement* and determine whether the student was an *underachiever, achiever,* or *overachiever* on a given subtest. One problem experienced when using the intellectual status index was that this technique tended to *overestimate* a given student's predicted achievement.

In 1969, Mykebust developed a *Learning Quotient* which is purported to be a more reliable estimate of expected performance. The learning quotient includes a more definitive and comprehensive index of expectancy as it takes into account not only the mental age, but the chronological age representing physiological maturity, and the grade age representing an index of school experience as well. A learning quotient can be computed for each area of achievement (each subtest) and therefore provides individualistic information of a diagnostic nature to help identify particular strengths and weaknesses for each child.

Actual achievement (scores earned on achievement tests) is then related to this expectancy and multiplied by 100 resulting in a Learning Quotient:

$$\frac{\text{Achievement Age} \times 100}{\text{Expectancy Age}} = \text{Learning Quotient}$$

Learning quotients of 89 and below are interpreted to indicate a substantial discrepancy between actual and expected achievement indicating a student learning deficiency requiring special attention. Learning quotients of roughly 90 to 93 represent problem areas still requiring attention but not quite as severe as the 89 and below L.Q.s. Learning quotients of roughly 94 and above indicate that a child is achieving at a level commensurate with his expectancy. A relative pattern of strengths and weaknesses becomes apparent as L.Q.s in different achievement areas are compared.

The Foundation For Individualized Evaluation and Research, Inc. (FIER) a non-profit foundation of DeKalb, Ill., provides low cost com-

puterized evaluation for school districts employing individualized approaches in their educational programs. FIER has been validated by the Westinghouse Learning Corporation to provide local evaluation of school districts utilizing Project PLAN.

Whether the local district employs a combination of norm and criterion-referenced data to assess their educational programs on a local basis, or embarks on a computerized approach to measure the individualized learning of its students, the results come out the same. Areas of program strengths and areas of improvement result. It remains then for the educational leaders in the district to evaluate these data, and initiate programs of change, where appropriate.

CHAPTER VI

STATEWIDE EDUCATIONAL EVALUATION

Some Technical Considerations

Some states have legislation which requires some form of state testing program. The obvious intent behind this legislation was to gather educational information which could provide the basis for further educational legislation.

The expression *statewide testing program* has fallen into misuse of late because of the emotional furor produced by the more vocal elements of the public when the word *testing* is associated with a county or state program. To counteract this public negativism, educators have tended to substitute the word *assessment* to describe their programs.

You are now faced with the prospect of having or not having some kinds of programs which will describe student learning at all grade levels in the state. These programs should be designed so local district curricula can be examined and revised if found wanting. At the state level, you need to know what kinds of student needs exist, and how you might best apportion the state educational resources to remediate these needs. You will also have to finance these projects from existing state educational resources, and therefore insist that the program cost be minimal. What do you do?

This dilemma is similar to one experienced by the former Dean of the Faculty of the U.S. Air Force Academy. He was concerned that the all-military teaching faculty had too much military rank and were getting older. He was attempting to minimize the costs associated with each cadet's education. The present faculty age and rank averaged 34 and major. He wanted a staff composed of rated pilots with over five years flying experience (combat preferred), having Ph.D. degrees in the areas to be taught at the Academy, exemplary military records, and he wanted the professors to be 25 years old and First Lieutenants! The same problem exists for you as state school board members as you decide to establish a program which will provide you with data reflecting the amount and quality of public education in your state. The funds you expend for this program will diminish the state educational resources available for distribution in support of needed or innovative educational programs at the local level.

Now, what should you call such a program? You could call it a *statewide testing program (STP)*, but the expression *testing* has a strong negative connotation. Some states have elected to call such a program a *Statewide Educational Assessment Program (SEAP)*. The expression assessment, being synonymous with measurement, does not reflect the

true purpose from such a program. Assessment, especially as used by National Assessment, merely describes the general conditions of learning without making comments relative to student learning in District X or District Y. For this reason, the expression assessment is an inappropriate rubric for the program.

A Statewide Educational Evaluation (SEE) program perhaps comes closer to describing what is desired than either of the other two suggested titles. (Note that the letter P was not included in the acronym, since *see* meaning *to perceive . . . . or examine* comes closer to the intent of the project, rather than *SEEP* meaning *a spot where water or petroleum oozes out slowly and gathers in a pool* (Webster's New Collegiate Dictionary). Literal elegance and accuracy aside, a Statewide Educational Evaluation program certainly represents what is desired. The program should measure or assess student learning by grade level. To accomplish this, some mix of norm and criterion-referenced instruments should be employed or developed. The results should then be compared to other state, regional or national data, and judgments relative to the quantity and quality of student learning should be undertaken (evaluation). Finally a summary of these findings must be developed so that educational leaders, teachers, legislators, parents, students and the public-at-large can read and understand them.

Frank Womer's monograph *Dereloping A Large Scale Assessment Program* describes in detail a systematic approach to develop and implement an assessment program. By experience and rigorous statistical training with National Assessment, Dr. Womer's paradigm very closely represents that model.

The procedures which follow represent a less rigorous but also a less costly alternative to developing, implementing and disseminating the results of a Statewide Educational Evaluation (SEE) program.

Developing the Program: The SEE program, to be a successful venture, must have the cooperative participation of legislators, state board of education members, the state department of education, educational leaders in the state, teachers, parents, students and the public-at-large. The planning stage is crucial, but it also must be managed within a reasonable (say 6 months) period of time. It has become an American educational practice to place possibly embarrassing programs in an indefinite tenure by forming a planning committee. This technique soothes the emotions of the program zealots while assuring that no immediate action will take place. After the committee makes its report, the recommendations are then retained in the local board of education's minutes thereby preserving its perpetual obscurity.

Such a planning mode is not envisioned for the SEE project. De-

veloping the SEE program will involve at least six major steps. These are (1) appointing a SEE Project advisory committee to the state board of education; (2) appointing an interim SEE Project director; (3) determining the objectives to be assessed by grade level and academic discipline; (4) developing a measurement and sampling strategy; (5) developing the instruments; and (6) pilot testing the instruments for validity and reliability.

SEE Project Advisory Committee: The state board of education should appoint a SEE Project advisory committee composed of representatives of state board of education, legislature, state educational professional societies, educators, the state department of education, business and commerce, labor, the trades, and students. The advisory committee must receive a specific charge from the state board of education which outlines: its duties; tenure (no more than 6 months); its role in coordination with the SEE Project director and his staff; and. the interval and requirements of the reports they should make to the state board of education. Meetings of the advisory committee will most likely progress on a decreasing frequency basis ranging from weekly to monthly during the tenure of its existence. The advisory committee approach allows all interested points of view to impact upon the embryonic program when modifications are relatively easy. The advisory committee should have a definite budget under which they carry out their activities. At its first meeting, the members will elect a chairman who will provide the leadership, schedule the frequency of meetings, and provide liaison to both the SEE Project director and the state board of education. The advisory committee's final report to the state board of education will summarize the activities, highlights, problems encountered, and recommendations.

One obvious recommendation could be that the tenure of the advisory committee should be extended for some duration. The implementation of that and other recommendations, however, then becomes the prerogative of the state board of education. Advisory committees will operate differently from state to state depending upon factors such as geographical size of the state, the funds available for travel and per diem, the need for and number of sub-committees which will be formed, etc.

Appointing an Interim SEE Project Director: Concomitant with the appointment of the SEE Project Advisory Committee, an interim SEE Project Director should be appointed. The interim SEE Project Director should be appointed for a two-year term and should possess an educational background rich in measurement, evaluation, and statistics. The interim SEE project director should be assigned to the state

department of education's staff so he reports directly to the chief state school officer. The interim SEE Project Director will manage all aspects of the SEE Project. He will select a staff, coordinate activities with other concerned agencies, supervise the sampling, instrument development, data collection and analysis, and the dissemination and follow up of the first final report. The director will be required to solve a myriad of problems in executing his term of office. These problems will be highlighted or eased depending on the size of his budget, the comprehensiveness of the sample, the use of norm-referenced or criterion-referenced instruments or whether he must supervise the development of a unique instrument, etc. After the first SEE report has been published and disseminated, both the state board of education and the interim director will probably want to take a *long look* at each other before the decision to support a permanent director is made.

The Educational Programs to be Measured: This particular planning component represents the key around which everything else must inevitably evolve. It is also this event which most probably would have the greatest variability from state to state. The notion that education is a state function is a widely recognized truism across the country. Some states reserve this prerogative and administer public education in a regulatory manner through the state department of education. Other states have chosen to have the local school district establish their own educational programs, and have enlisted the state department of education personnel to function in a service role to local educators. Still other states employ some combination of these two extremes in administering the educational programs in their state.

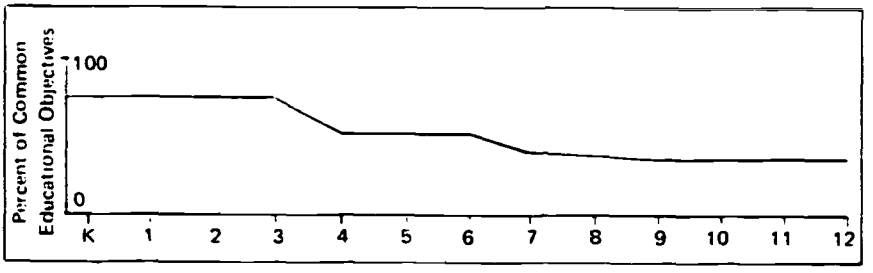
The statement of common state goals and objectives for education will vary depending upon the leadership style employed at the state level. Regardless of the leadership style, there would be some common as well as different educational programs which exist at each of the K-12 grade levels. These common elements are hypothetically displayed in Figure 4.

The hypothetical information on common educational programs presented in Figure 4 is based upon no data, but experience that suggests in grades K-3 approximately 80% of the in-school time is devoted to student learning of basic skills (reading, language arts, and math). During the upper elementary, junior high, and high school years the number of common objectives from one school district to another would be more likely to decrease. The SEE Program could only validly measure those objectives commonly shared throughout the

FIGURE 4

The Percentage of Common Educational Program Objectives

(Hypothetical Data)



state. Local educational evaluation is then required to bridge the gap between the state measurement of common objectives and the total number of objectives in their local programs.

It would be estimated that teams of educators representative of the K-12 academic disciplines would need to be convened to screen out the multitude of local objectives and to rephrase the common objectives so they can be classified (according to Bloom), and measured later by questions on the statewide evaluation. This list of measurable objectives then should be coordinated with the state department of education, and local school district curriculum specialists before being presented to the state board of education for endorsement.

During this period, decisions regarding the number of grade levels and academic disciplines to be measured must be made. The funding levels available for the SEE Program will in all likelihood cause a compromise from an ideal distribution of grade levels and subject matter to be measured to a more real level of effort.

Develop a Measurement and Sampling Strategy. The 1973 Educational Testing Service report *State Educational Assessment Program* indicates testing programs in existence in the 50 states and territories. These programs vary depending on the sources of funding (Titles I and III ESEA, state statute for reading, state department of education programs, etc.) Womer (1973) suggests the techniques of matrix sampling employed by National Assessment as an efficient technique. Matrix sampling is *like selecting a series of separate, non-overlapping samples of students, with each sample taking a different subset of the total number of items to be administered* (Womer, 1973, p. 70).

This technique has much to recommend it. Instead of asking each student to devote 3 hours to take the SEE, three students could take non-overlapping items in one hour. While this technique increases the total number of students to be involved in the project, it does so with minimum interruption of the school program.

Some states may choose to measure every student in specific grades and gather information about each child three or more times during the 13 years the child spends in the system. Figure 5 describes a variable three-year cycle to evaluate students at each grade level once every three years.

FIGURE 5

Three Year Evaluation Plan

Year I	Year II	Year III
1	1	(1)
2	(2)	2
(3)	3	3
4	4	(4)
5	(5)	5
(6)	6	6
7	7	(7)
8	(8)	8
(9)	9	9
10	10	(10)
11	(11)	11
(12)	12	12

The three-year evaluation plan described in Figure 5 allows two intervening years to revise the educational program for a given grade level before that grade is measured again. This three-year evaluation plan would be more feasible at the local level (providing they used norm-referenced tests) than at the state level. If this three-year cycle were employed to develop four unique instruments each year, the cost of such an undertaking would be prohibitive.

The decisions to go with sampling versus *every pupil* techniques, and whether the state elects to measure students at one or more grade levels need to be made early in the planning state. Variation in the measurement and sampling strategy later in the enterprise cause not only considerable adjustments to be made, but also additional funds to defray these changes.

Develop the Instruments: Test development is a very tedious, time consuming activity. Major testing companies employ large numbers of test item writers on their staff, as well as college and university faculty consultants. The interim director of SEE will have to devote considerable care and attention to selecting appropriate staff members for the purpose. He may decide to contract this portion of the program to professional testing companies whose staff has this expertise, but this approach is expensive.

Should he decide to attempt to generate his measuring instruments within the expertise of state, local and university agencies, he has a few procedural techniques which will tend to make the job easier. You may recall in the section dealing with learning, *The Taxonomy of Educational Objectives* (Bloom and Krathwohl editions), was used to describe the levels of learning in the cognitive and affective domains. Using this taxonomy, the objectives to be measured can be classified relative to the level of learning. The taxonomy also contains sample test item formats which the authors purport are designed to measure a given objective at the appropriate level. Thus, the intelligent use of the taxonomy to develop questions to measure objectives at a given level of learning results in an additional criterion standard with which you can provide content validity for the individual test items.

Womer cites the fact that National Assessment willingly makes test items available to educators at no cost. The items, the objectives they were designed to measure, and the ease indices (proportion of students who scored correctly on the item) are made available by National Assessment. It must be assumed that for the many assessment instruments developed by National Assessment, a large number of test items (with demonstrated item analysis) would be appropriate for the SEE instrument. The need then exists to generate test items which would measure all of the remaining objectives at the appropriate level of learning.

To accomplish the writing of the test items, it is recommended that a team of teachers be identified for a given academic discipline and grade level. It is further recommended that an experienced test writer or educational psychologist (with a background in psychometrics) also be assigned to the team. It is his task to: (1) explain the classification and test item rationale of the taxonomy to the team; (2) advise them on the general guidelines for effective writing of test items (no trick questions, no double negatives, don't ask a series of questions where if the student can get one right, he can through elimination answer the others, don't use the words *always* or *never* in true-false questions because they invariably point to a false response, etc.);

and, (3) serve as an evaluation consultant to the team. The team members after the above orientation will decide the specific kinds and types of questions to ask as well as being responsible to assure that the specific vocabulary used in the question is compatible to the reading level of the typical student in the grade measured. Affective as well as cognitive questions should be contained in the SEE instrument.

A separate team would be required to generate questions for each academic discipline at each grade level. The items generated by a given team should be reviewed (1) by other qualified teachers for that grade level and subjects, and (2) by the project staff.

The Pilot Study: When the SEE instrument is developed, it should be pilot tested on a small group to assure that the items are valid and that the subtests are reliable. The size of the pilot sample should be a few hundred students who are not to be included in the larger statewide sample.

When the results are available, the test should undergo a thorough item analysis. The item analysis will reveal the ease index for each item, the specific responses which are not being selected and the estimate of reliability. Traditional testing protocols would insist that each subtest start with easy items and that the subtest get more difficult as the student progresses. If the plan were to follow a norm-referenced approach, the items could simply be ordered in terms of increasing difficulty as a result of the item analysis. National Assessment has found that the random presentation of items by difficulty can often result in student reinforcement because as he goes along he may encounter an easier question after a few hard ones, and the reward of the easy question challenges him to continue. Each state must decide for itself how it wishes to handle the ordering of items on its SEE.

Implementing the Program: Once the sample and the measuring instrument have been developed and refined the interim director of SEE can breathe just a little easier. The decision as to when the testing would take place should already have been anticipated, discussed, and finalized. Some states employ fall testing so the results can be analyzed and reported back to the districts while the children are enrolled in the same academic year. This practice has much to recommend it. If the local educators find that some students in their district experienced difficulty with certain concepts, these concepts could be retaught before the child is promoted to the next higher grade.

Basically there are two crucial components to the implementation of the SEE program. These are: distribute the instruments; and collect, score and analyze the data.

Distribution of the Instruments: As part of the sampling procedure, the SEE project staff should have informed the school districts selected for inclusion in the project of the number of students participating and the projected testing dates. The final instruments and answer sheets should be packaged and mailed to the school districts so they arrive in the local district at least one week prior to the desired testing date.

Each district should have appointed a SEE project teacher (or administrator) who would handle the logistics of explaining the testing instructions to local district test monitors and assuring that each monitor checked both test booklets and answer sheets at the close of the testing session. The local SEE project teacher would then mail the answer sheets and test booklets (under separate cover) to the state SEE project staff.

Data Analysis: When the answer sheets arrive at the SEE project office, they are checked for stray marks, multiple answers, etc. and prepared for some form of machine scoring. The faster procedure would be to have the answer sheets designed so they could be run through optical scanning input devices for computers. Of course, alternative scoring procedures could be employed, but they tend to be more time consuming. The data on each answer sheet is then prepared for computer processing; and, when all the answer sheets are available, the first of a series of computer runs is undertaken.

The first run through the computer should merely tabulate the frequency count for each item on the test. With these data, the reliability of the test can be readily ascertained. Comparison of the pilot-test data and the actual sample then can be made and differences noted.

Further runs might find the total sample partitioned on some variable, i.e., sex, ethnic background, socio-economic status, district size, college bound youngsters, vocational education students, etc. The item analyses generated from these additional runs might reveal areas of educational need, program strengths, etc.

Specific computer runs to group the item data by educational objective would also be undertaken. The print-outs then would be analyzed and compared between the total group and each of the variables listed above.

Additional statistical analysis such as factor analysis to see whether the instrument actually measured just what you thought it was measuring or whether additional factors emerge from the analysis should also be undertaken.

Norms then would be developed on a total state, district size and/or geographical region bases so local educators can appraise their results in context with other similar districts.

Throughout the analyses, the anonymity of each school and district must be scrupulously adhered to.

When all the analyses have been completed, percentiles should be derived for norm-referenced instruments as well as learning quotients if they are desired. For norm-referenced measurement the average ease index for each objective must be obtained, and a decision of whether the students met the specific criterion must be established. These data should then be displayed in summary fashion for use in the SEE report.

Disseminating the SEE Project Reports

Dissemination:

The reports generated by the SEE project must be written in different forms so the targeted audience can read and understand them. A technical report replete with statistical comparisons can be developed for educational researchers and key educational administrators. An "English" version of the report should be developed so parents, students and the public-at-large can read and react to it.

The outline of each version of the SEE program report should follow the taxonomy for educational decision-makers described in the introduction of this handbook. The SEE evaluation data should provide partial answers to:

1. *What should our school graduates: know; be able to do; believe? What do they know?*
2. *To what extent: should the schools ameliorate class differences; emphasize individual excellence; insist on common minimum standards of performance; treat the exceptional student exceptionally?*
3. *What do our citizens believe: should be the goals of education; are the essential priorities of education; about the value of student preferences and needs?*
4. *What should we do about the contradictions affecting education: between freedom, progress and necessity; between goals of equality and goals of excellence; between what we believe and what we are doing; and between societal and individual needs and preferences?*

If the SEE Program reports respond to this taxonomy of educational questions, the achievements, the successes, the areas for improvement, and the follow up activities come sharply and quickly into focus. The report should provide the locus for innovation and improvement in public education and allow you, the state board of education member, to receive *hard data* so you can make more valid educational decisions for the future.

CHAPTER VII

SOME FINAL CONSIDERATIONS

This handbook would not be complete without comment on some global considerations important to state board of education policy-making with respect to statewide educational evaluation. These comments may be viewed also as thinly disguised advice.

The state board of education establishes policy; the chief state school officer and the state education agency staff execute that policy. This basic and oft-repeated tenet is doubly important when it comes to assessment. State board members should find out all they want to know about the programs. They should set the policy limits and general timetable, and then delegate policy execution to the staff. The state board of education should call for frequent progress reports and expect results, but state board members need not dabble in the day-to-day operational details. Attention to the broad overall objectives will be sufficient. This will not only save a lot of board time, but the staff will respect the confidence which has been placed in them.

It is not essential that a state have a statewide educational assessment program in order to have good schools - but an assessment program may help in showing ways to improve those good schools. For an assessment program is only one major school management technique having potential for effectiveness, it is not a panacea. A poorly conceived, underfinanced, and badly executed assessment program is worse than no assessment program at all. Rather than have a poor program, a state should perhaps look for current activities and programs which can become components of an assessment program and build from there. For example, almost every state education agency collects much information from local districts that is never used or analyzed at the state level; a bit of time and money spent sifting through this for indications of school performance might return bigger dividends in decision making information than time and money spent in starting a new assessment program.

And speaking of *decision-making*, let's take a frank look at what it really means. Decision making information might mean to the educator those facts which substantiate the need for additional state funds being provided to the schools. Decision making information to the legislator might mean those facts that show high-cost-per-student districts have students whose performance is less on the average than districts which have low per pupil costs. An assessment program whose objectives are set to confirm predesignated opinions like these may be destined to failure.

Proponents of statewide educational evaluation apparently believe that additional *hard* data, provided through scientific methods and presented in an objective manner, will suddenly elevate decision-making to a high-level, completely rational activity. It may help in this regard, but let's face it: decision-making is still accomplished by people, and people will still make decisions on how they *feel* about things. The best assessment information in the world will not change this.

A statewide educational evaluation program which first concentrates on information for decision making at the local level probably stands a better chance of initial acceptance and long-term success. The choice of first serving local districts — administrators and boards will indicate state board recognition that the most important decisions affecting a child's education are made at the local level. Those states which have given priority in recent years to the development of management information systems at the local level will find this investment paying off since assessment procedures can merely become an *add-on* to current programs.

It is important here to review the position on assessment taken by chief state school officers in 1971. The policy statement was as follows:

> *The Council of Chief State School Officers recognizes the necessity for the assessment of education at the state and local level and urges the member states to support the development of assessment capability within the states.*
>
> *The Council reaffirms its commitment to a fuller understanding of the status and needs of American public education through educational assessment at the national level as conducted by the Education Commission of the States.*
>
> (State & Federal Relationship in Education)

Perhaps the very first thing to do when considering an assessment program is to define what it would mean in a particular state. Also, define the related terms, like needs assessment, accountability, state testing program, learner needs assessment, student performance assessment, accreditation, district assessment, and student assessment. Development of a glossary of terms containing these and other words will contribute to understanding as discussions proceed to: Shall we assess? Why assess? and What shall be assessed?

Is educational assessment here to stay? Certainly, but to put this answer in perspective, remember that: (1) state boards of education and state education agencies have been charged with reporting the status of public education since their establishment; (2) the elements of an assessment program have been around for quite a while; it's the packaging of these elements to provide certain kinds of information

that's new; and, (3) some new-sounding educational management technique will be along in a year or so and overshadow the current interest in assessment.

So examine assessment, use the idea and measurement techniques in those situations where it is warranted, and don't expect it to be the ultimate answer to anything. The admonition here to those considering assessment is to suggest one ultimate question to be answered: *Will what we are doing help kids?*

In this little volume *assessment* has been presented synonymously with *measurement*. Statewide assessment has its advocates, its advantages and its dangers. Taken by itself assessment without evaluation is sterile, and if it doesn't lead to good decisions and workable decisions, it may be anything from useless to evil. Assessment can not intelligently be used to judge the quality of individual teachers. The factors of learning will not permit such naive simplicity. Some will nevertheless advocate such nonsense.

A greater evil by far would be using assessment data to justify arbitrary standards for quality education which will early in school careers classify and categorize opportunities for individuals. The schools of the United States flowed from our democratic orientation and basic disdain for class structure. We must not succumb to the temptation to use education to stratify our people. We mustn't abandon the optimism and hope, perhaps the naivete, of a system open to all people for all their lives including the democratic right to fail in programs too difficult.

Lastly let us approach statewide evaluation with caution. Let us not make glorious promises to our citizens which will fall short and disappoint. We don't want statewide evaluation to become another dead fad like progressive education, teaching machines, programmed learning, flexible scheduling and other gimmicks oversold to the public by zealous and gullible educators.

REFERENCES

Ahmann, J. Stanley *The First Results* in **Compact**. Denver, Colorado Education Commission of the States - Vol. 6, No. 1, February 72 (pp. 13-17)

Anastasi, Anne *Psychological Testing* - New York - The Macmillan Co. - 1961

Bloom, Benjamin S. (Ed.) *Taxonomy of Educational Objectives Handbook I: Cognitive Domain* - New York - David McKay Co. Inc. 1956 (pp. 201-207)

_____ *State Educational Assessment Programs* - Princeton, N.J. Educational Testing Service - 1971

_____ *State Educational Assessment Programs* - 1973 Revision - Princeton, N.J. Educational Testing Service - 1973

Bruno, Nancy L., Paul B. Campbell, and William H. Schabacker *Statewide Assessment: Methods and Concerns* - Princeton, N.J.: Center for Statewide Educational Assessment, Education Testing Service - 1972 (25 pp.)

Compact - Education Commission of the States, Vol. 6, No. 1, Feb. 1972 (issue on National Assessment: Measuring American Education)

Gardner, John William *Excellence, Can We Be Equal and Excellent Too?* - New York, Harper - 1961 (171 pp.)

Jencks, Christoper *Inequality: A Reassessment of the Effect of Family and Schooling in America* - New York - Basic Books - 1972 (399 pp.)

Krathwohl, David R. (Ed.) *Taxonomy of Educational Objectives Handbook II: Affective Domain* - New York - David McKay Co. Inc. - 1965 (pp. 176-185)

Implications for Education of Prospective Changes in Society - Denver: Designing Education for the Future - 1967 (p. 195)

Myklebast, H.E. (Ed.) *Progress in Learning Disabilities* - Vol. 1 - New York - Greene Stratton - 1969

Nolte, M. Chester *An Introduction to School Administration: Selected Readings* - New York - The Macmillan Company - 1966 (p. 74)

Olson, Arthur R. and Edward H. Lyell *Educational Evaluation and Assessment System for You* - Denver, Colo. Department of Education - 1972 (48 pp. plus appendices)

Phi Delta Kappan, Vol. VII, No. 4 - Dec. 1970 (8 articles on accountability)

Phi Delta Kappan, Vol. LIV, No. 1 - Sept. 1972

The Future as Metaphor: Educational Decision-Making - Seminar Report: Regional Interstate Project Program - Denver, Colo. Dept. of Education - 1972 (67 pp.)

Womer, Frank B. *Developing a Large Scale Assessment Program* - Denver, Colo. - Cooperative Accountability Project Monograph - 1973

APPENDIX I
DIRECTORY OF KEY SEA ASSESSMENT AND EVALUATION PERSONNEL

ALABAMA

Ledford L. Boone, Coordinator
Planning & Evaluation
State Department of Education
State Office Building
Montgomery, Alabama 36104

ALASKA

Ernest E. Polley, Coordinator
Planning & Research
State Department of Education
Pouch F, Juneau, Alaska 99801

AMERICAN SAMOA

Tyman L. Stephens
Asst. Director Business Services
Department of Education
Pago Pago, American Samoa 96799

ARIZONA

William R. Raymond, Director
Planning and Evaluation
1535 W. Jefferson
Phoenix, Arizona 85007

ARKANSAS

Dr. Sherman Peterson, Associate Director, Planning & Evaluation
Department of Education
Education Building
Little Rock, Arkansas 72201

CALIFORNIA

William Bronson
Office of Program Evaluation
State Department of Education
721 Capitol Mall
Sacramento, California 95814

CANAL ZONE

Dr. C. L. Latimer
Deputy Superintendent of Schools
Box 123
Balboa, Canal Zone

COLORADO

J. D. Hennes, Consultant
Planning and Evaluation
Department of Education
Denver, Colorado 80203

CONNECTICUT

R. Douglas Dopp
Special Projects Planning
Room 310 State Office Building
Connecticut State Department of Education
Hartford, Connecticut 06115

DELAWARE

Robert A. Biglow
Planning, Research & Evaluation Division
State Dept. of Public Instruction
Dover, Delaware 19901

DISTRICT OF COLUMBIA

Robert B. Farr, Director
Pupil Appraisal
415 12th Street, N.W.
Washington, D.C.

FLORIDA

James C. Impara
Administrative Education Acct.
Department of Education
Tallahassee, Florida

GEORGIA

Sarah H. Moore
Coordinator for Evaluation
State Office Building
Room 315 Annex
Atlanta, Georgia 30303

GUAM

James B. Branch
Administrative Head
Planning & Evaluation Unit
Department of Education
P.O. Box DE
Agana, Guam 96910

HAWAII

Ronald L. Johnson, Administrator
Evaluation Section - OIS
Department of Education
1270 Queen Emma Street
Honolulu, Hawaii 96813

IDAHO

Wayne A. Phillips, Prog. Admin.
Planning, Development & Info.
State Department of Education
Len B. Jordan Building
Boise, Idaho 83720

ILLINOIS

Thomas Springer, Director
Statewide Assessment &
Evaluation Section
Office of Supt. of Pub. Instruction
325 South 5th Street
Springfield, Illinois

INDIANA

Ivan Wagner, Director
Planning and Evaluation
State Office Building
Room 803
Indianapolis, Indiana

IOWA

Max Morrison, Director
Planning, Res. & Evaluation
Dept. of Public Instruction
Grimes Office Building
Des Moines, Iowa 50319

KANSAS

Dr. Larry Casto, Asst. Comm.
Division of Development
State Department of Education
120 E. Hampden Street
Topeka, Kansas 66612

KENTUCKY

Donald S. Van Fleet
Director of Evaluation
State Department of Education
Frankfort, Kentucky 40601

LOUISIANA

Katherine P. Finley, Director
Planning and Evaluation
Deputy Assoc. Supt.
of School Programs
P.O. Box 44064
State Department of Education
Baton Rouge, Louisiana 70804

MAINE

Dr. Horace P. Maxcy, Jr.
Planner
Education Building
Augusta, Maine 04330

MARYLAND

Richard K. McKay
Asst. State Superintendent
State Dept. of Education
P.O. Box 8717
Friendship International Airport
Baltimore, Maryland 21240

MASSACHUSETTS

James F. Baker
Associate Commissioner
182 Tremont Street
Boston, Massachusetts 02111

MICHIGAN

Robert J. Huyser
Supervisor, Assessment Program
Box 420
Lansing, Michigan 48902

MINNESOTA

John W. Adams, Director
State Education Assessment
731 Capitol Square Building
St. Paul, Minnesota 55101

MISSISSIPPI

Jerry R. Hutchinson, Coor.
Office of Planning & Evaluation
Post Office Box 771
Jackson, Mississippi 39205

MISSOURI

John F. Allan, Director
Planning & Evaluation
Missouri State Dept. of Education
Jefferson City, Missouri 65101

MONTANA

Mike Pichette
Reporting Services Coordinator
Office of Supt. of Pub. Instruction
Helena, Montana 59601

NEBRASKA

Francis E. Colgan, Administrator
Planning, Evaluation & Research
State Department of Education
233 So. 10th Street
Lincoln, Nebraska 68508

NEVADA

James Kiley, Assoc. Supt.
Division of Planning & Evaluation
Nevada State Dept. of Education
Carson City, Nevada 89701

NEW HAMPSHIRE

R. Schweiker, Senior Consultant
Research and Testing Services
N. H. Department of Education
Concord, New Hampshire 03301

NEW JERSEY

Bernard A. Kaplan, Director
Our Schools Prog. Division of Research, Planning and Evaluation
State Department of Education
224 W. State Street
Trenton, New Jersey 06825

NEW MEXICO

Alan Morgan, State Director
Evaluation and Assessment
Education Building
Santa Fe, New Mexico 87501

NEW YORK

L. Woollatt, Assoc. Commissioner
N. Y. State Dept. of Education
Albany, New York 12224

NORTH CAROLINA

William J. Brown, Director
Division of Research
N. C. Dept. of Public Instruction
Raleigh, North Carolina 27611

NORTH DAKOTA

Lowell L. Jensen, Director
Div. of Planning & Devel.
State Dept. of Public Instruction
Bismarck, North Dakota 58501

OHIO

Roger J. Lulow, Director
Div. of Planning & Evaluation
Room 615, 65 So. Front Street
Columbus, Ohio 43215

OKLAHOMA

James Casey, Coordinator
Planning, Research, Evaluation
State Department of Education
State Capitol
Oklahoma City, Okla. 73105

OREGON

R. B. Clemmer, Coordinator
Planning & Evaluation
942 Lancaster Drive NE
Salem, Oregon 97310

PENNSYLVANIA

Thomas E. Kendig, Chief
Division of Educational Quality Assessment
Pennsylvania Dept. of Education
P.O. Box 911
Harrisburg, Pennsylvania 17126

PUERTO RICO

Marta Barros Loubriel, Acting Dir., Evaluation Program
Urb. Tres. Monitas-Calle Teniente Cesar Gonzales esq. Calaf.
Hato Rey, Puerto Rico 00919

RHODE ISLAND

Cynthia V. L. Ward
Education Research Specialist
Department of Education
Division of Research and
Evaluation, Room 210
Hayes Street
Providence, Rhode Island

SOUTH CAROLINA

Dr. W. E. Ellis, Director
Office of Research
S. C. Department of Education
Columbia, South Carolina 29201

SOUTH DAKOTA

Dr. Henry Kosters, Asst. Supt.
Division of Elem. & Secon. Educ.
Capitol Building
Pierre, South Dakota 57501

TENNESSEE

John N. Hooker
Director, Testing Sevrices
V.T. at Knoxville
1000 White Avenue
Knoxville, Tennessee 37916

TEXAS

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

TRUST TERRITORY

Thomas R. Brown
Program & Research Officer
Office of the High Commissioner
Trust Territory of the
Pacific Islands
Saipan, Marianna Islands 96950

UTAH

Stephen L. Murray
Evaluation Specialist
1400 University Club Building
Salt Lake City, Utah

VERMONT

Dr. Herbert Tilley, Director
Planning & Evaluation
State Office Building
Department of Education
Montpelier, Vermont

VIRGINIA

Dr. Charles C. Todd, Jr.
Director of Planning
State Department of Education
Richmond, Virginia 23216

VIRGIN ISLANDS

Peter Rasmussen, Director
Planning, Research & Evaluation
Box 630, Dept. of Education
Charlotte Amalie, St. Thomas
Virgin Islands 00801

WASHINGTON

Alfred Rasp, Jr.
Director, Program Evaluation
Old Capitol Building
Olympia, Washington 98504

WEST VIRGINIA

Dr. B.G. Pauley, Asst. Supt.
Bureau of Services & Federal Prog.
Department of Education
Capitol Complex, Bldg. 6-B
Charleston, West Virginia 25305

WISCONSIN

James H. Gold
Division for Planning Services
Dept. of Public Instruction
126 Langdon Street
Madison, Wisconsin 53702

WYOMING

Paul D. Sandifer, Asst. Supt.
for Planning, Evaluation &
Information Services
State Department of Education
Capitol Building
Cheyenne, Wyoming 82001

APPENDIX II

CONTRACTING AGENCIES USED BY STATE EDUCATION DEPARTMENTS* FOR MATTERS RELATING TO ASSESSMENT

Alabama	
Needs Assessment	College of Education, University of Alabama
Alaska	
Needs assessment	Northwest Regional Education Laboratory, Portland, Oregon Stanford Research Institute Brookings Institution
Arizona	
Analysis of data for reading achievement test	Southwest Research Associates
Needs assessment	Arizona State University
Needs assessment	EPIC Diversified Systems, Corp., Tucson, Arizona
Needs assessment	Consulting Services Corp., Seattle, Washington
Arkansas	
Needs assessment	EPIC Diversified Systems, Corp., Tucson, Arizona
California	
Scoring services for statewide testing program	California Test Bureau/McGraw Hill, Monterey, Calif. School Testing Service, Berkeley, Calif.
PPBS	Peat, Marwick, Mitchell & Co. (management consulting firm)
Colorado	
Assessment of learner needs (edited assessment exercises)	University of Colorado, Laboratory of Educational Research
Assessment of learner needs (computer programming consulting)	Pacific Educational Evaluation Systems (PEES)
Assessment of learner needs (specifications and computer programs to analyze responses)	Automated Data Processing Service Center, State of Colorado
Assessment of learner needs (exercises to assess affective learning)	Interstate Service Center
Connecticut	
Needs assessment and educational goals	Institute for the Study of Inquiring Systems (ISIS), Philadelphia

**This information was gathered under the sponsorship of the Cooperative Accountability Project, Denver, Colo. Arthur R. Olson, Director.*

Delaware	
Needs assessment	Educational Testing Service, Princeton, N.J.
District of Columbia	
Reading and mathematics testing	California Test Bureau McGraw Hill
Florida	
Statewide assessment program, instruments and objectives	Center for the Study of Evaluation (CSE), University of California at Los Angeles
Statewide assessment program, test scoring	Software Programming and Associates, Titusville, Florida
Catalogue of reading objectives	Florida State University
Illinois	
Needs Assessment (tests and scoring devices)	Science Research Associates, Chicago, Illinois
Iowa	
Needs assessment (criterion-referenced measures)	Instructional Objectives Exchange, UCLA
Kansas	
Needs assessment	Research and Grants Center, Kansas State Teachers College, Emporia
PPBS	Mid-Continent Regional Educational Laboratory, Kansas City, Mo.
Evaluation of vocational education	Teaching Research Division, Oregon State System of Higher Education
Kentucky	
Needs assessment	EPIC Diversified Systems Corp.
Louisiana	
Needs assessment	Research Division, Northwestern State College of Louisiana
Maine	
Statewide assessment	Research Consortium for Educational Assessment (includes Research Triangle Institute, N.C.; Measurement Research Center, Iowa City, Iowa; and the American Institute for Research, Palo Alto, California.)
Maryland	
Instructional evaluation studies	Institute of Administrative Research, Teachers College, Columbia University
Goals	Automation Industries, Inc., Vitro Laboratories Division

Massachusetts	
Fourth grade testing program	Educational Testing Service, Princeton, N.J.; CBT/McGraw Hill; Project Comprehensive Achievement Monitoring; Instructional Objectives Exchange, UCLA; Westinghouse Learning and Measurement Research Center, its subsidiary.
Michigan	
Michigan Educational Assessment Program	Educational Testing Service, Princeton, N.J.
Minnesota	
Minnesota Educational Assessment Program	Research Triangle Institute, N.C.; Educational Management Services, Inc. University of Minnesota (for consulting)
Missouri	
Assessment of 4th & 6th grade skills	CTB/McGraw Hill
Statewide assessment program	Possibly Center for Education Assessment, Princeton, N.J.
Montana	
Educational needs study	Arthur D. Little, Inc., Cambridge, Mass.
Nevada	
Management information system	Dahl/Kramer. Project Consultants
Needs assessment, affective objectives and measures	Instructional Objectives Exchange, UCLA
New Jersey	
Goals	Opinion Research Corp., Princeton, N.J.
Statewide assessment	Educational Testing Service, Princeton, N.J.
New Mexico	
Needs assessment (does not look as though any contracting will be done for N.M.'s 1973 assessment, however	EPIC Diversified Systems Corp., Tucson, Arizona
New York	
Accountability system for the N.Y.C. school system	Educational Testing Service, Princeton, N.J.
Performance Indicators in Education Program	University of the State of New York
North Carolina	
State Assessment of Educational Progress	Research Triangle, N.C.

Ohio	
Development of educational accountability model	Ohio State University's Evaluation Center; president of Ohio Council for Education employed on a contractual basis.
Oklahoma	
Goals	College of Education, Oklahoma State University
Needs assessment	College of Education, Oklahoma State University
Oregon	
Oregon Assessment of Educational Progress	Science Research Associates, Chicago, Ill.
Pennsylvania	
Educational Quality Assessment	Educational Testing Service, Princeton, N.J.; Pennsylvania State University
Rhode Island	
PPBS	Peat, Marwick, Mitchell & Co. (conducted seminar on PPBS sponsored by Department of Education)
South Dakota	
PPBS	Applied Management Corporation
Tennessee	
Needs assessment	Memphis State University, College of Education
Texas	
Needs assessment - 6th grade reading - criterion-referenced instrument - also for 6th grade mathematics	CTB/McGraw Hill
Virginia	
Needs assessment	Bureau of Educational Research, University of Virginia
Washington	
Needs assessment	Consulting Services Corporation, Seattle, Washington
West Virginia	
Needs assessment - objectives, identifying variables	Human Resources Institute, West Virginia University

Wyoming

Needs assessment and goals	Center for Research, Service, and Publication, College of Education, University of Wyoming

Other agencies and institutions which state boards might wish to consult on evaluation matters include:

The American College Testing Program, P.O. Box 168, Iowa City, Iowa 52240
Foundation for Individualized Evaluation and Research, De Kalb, Illinois
General Learning Corporation, New York, New York
Education Commission of the States, Lincoln Tower, Denver, Colorado 80203
Grady Research Associates, 4604 El Camino Dr., Colorado Springs, Colo. 80918
National Assessment of Educational Progress, Denver, Colorado
American Council on Education, 1 Dupont Circle, Washington, D.C. 20036

Note: Any listing here does not imply NASBE endorsement or recommendation. State boards are advised to check on any consulting organization before contracts are let.