DOCUMENT RESUME

ED 097 366 95 TM 004 001

AUTHOR	Mushkin, Selma J.
TITLE	A Proposal for a "SIR" Adjusted Index of Educational Competence.
INSTITUTION	Georgetown Univ., Washington, D.C. Public Services Lab.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	DHEW-OE-74-11112
PUB DATE	Aug 73
CONTRACT	OEC-0-70-4454
NOTE	60p.
AVAILABLE FROM	Superintendent of Documents, U. S. Government Printing Office, Washington, D.C. 20402 ($1.00)
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	*Academic Achievement; Achievement Tests; *Comparative Analysis; *Demography; Education; Equated Scores; *Evaluation Methods; Income; Policy Formation; Program Effectiveness; Race; Research Methodology; Research Problems; Sex Differences; *Testing Problems; Testing Programs; Test Results; Universal Education
IDENTIFIERS	Educational Outcomes; *SIR Adjusted Index

ABSTRACT

The increasing use of educational performance or outcome measurements for a range of policy purposes points to new procedures for adjusting data for population composition. The purposes include: program formulation, budget resource allocation, grant-in-aid designs, performance incentive payments, consumer information for school selection, and program evaluation and review. This paper outlines methods for controlling population differences to make data on performance more comparable across time and from place to place. The resulting estimates of achievement scores, standardized for population differences, are useful for comparison only. Such comparative indexes remove the influence on average scores of population changes over time, or population differences between schools or school districts. Adjusted scores are not intended to take the place of the basic data but to complement them. Standardization procedures can be applied to achievement test scores and to other measurements of competence such as attitudes or attributes. In this report, achievement score adjustment is used as an example. The selection of sex, income, and race (SIR) as control variables is proposed as a first step. (Author/SE)

DHEW Publication No. (OE) 74-11112

A PROPOSAL FOR A "SIR" ADJUSTED INDEX OF EDUCATIONAL COMPETENCE

AUGUST 1973

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Caspar W. Weinberger, Secretary
Charles B. Saunders, Jr., Acting Assistant Secretary for Education

Office of Education
John Ottina, Commissioner

This report was prepared by Selma J. Mushkin, Director, Public Service Laboratory, Georgetown University, Washington, D.C., pursuant to Contract No. OEC-0-70-4454 with the Office of Education, U.S. Department of Health, Education, and Welfare.

Opinions expressed in the report are those of the author and do not necessarily represent Office of Education position or pɔlicy.

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1974

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 - Price $1

FOREWORD

Data on educational outcomes are increasingly being sought to better understand the results of resource commitment to education and to provide information for parents, teachers, and administrators in the quality of educational services. Has the output of schools increased over time? Are more children learning and learning more? Is educational performance better in one community than another, in one school than another?

Comparisons traditionally made between schools and between school districts could be improved by borrowing from the demographer's toolkit the notion of a reweighted arithmetic average. Standardization to compare population characteristics not under the schools' control is the kernel of Dr. Mushkin's proposal, which is still, we realize, in an embryonic stage. The adjustments are only for group scores, such as averages, not for individual students' scores.

Other methods, such as that developed by Henry Dyer for New York City schools, are available for refining school achievement comparisons.

The present report is distributed for comment or review as a part of an exploratory study on educational outcomes supported by the National Center for Educational Statistics. Public Services Laboratory of Georgetown University proposed the initial draft of this report in 1971 under contract with the U.S. Office of Education.

William Dorfman, Chief
Statistical Systems Branch

Dorothy M. Gilford
Assistant Commissioner
for Educational Statistics

ACKNOWLEDGMENTS

A large debt is owed to those busy persons who gave so willirgly of their time in serving as members of an ad hoc committee established to consider the "SIR" adjustments. It is hard to convey the author's deep pleasure and gratitude for the opportunity to work with this committee, whose members included: Alfred Carlson, Education Testing Service; William Coffman, Iowa Testing Programs; H. Russell Cort, General Learning Corporation; Burton R. Fisher, University of Wisconsin; Nelson Noggle, Science Research Associates.

To my colleagues at Georgetown, Henry S. Shryock and Jacob Siegel, far more knowledgeable than the author about standardization of death rates for differences in age of population, go a special thanks for their patience as well as expertise. Any errors, the author hastens to add, are her own and do not reflect on the valuable expert aid given.

Research assistants at Public Services Laboratory (PSL) helped put together materials on testing. Their assistance has greatly improved the document. Work was carried out over an extended period and, if some students who aided are not mentioned, the author hopes they will understand. A special "thank you" goes to David Bain, Steve Stageberg, and Ellen Weissbrodt.

Staff of the U.S. Office of Education participating in the ad hoc committee included Dorothy Gilford, Boyd Ladd, Ezra Glaser, Richard Berry, and William Dorfman; they contributed greatly to the usefulness of the discussions and to the cooperative spirit that all who engaged in this committee's work will long remember.

Selma J. Mushkin
Director, Public Services Laboratory
Georgetown University

CONTENTS

Page

Foreword..iii

Acknowledgments...v

Summary...ix

Introduction..1

Purposes of Interjurisdictional and Intertemporal Comparison................4

I school systems assuring equality of opportunity?..........5
What do children learn?..........6
How are the schools to be held accountable for their performance?..........6
What is education's role in social accounting?..........8

Status of Achievement Testing in State and Community..........10

What portion of school children are now routinely tested?..........11
How many different tests are given?..........12
At what grade levels are the achievement tests given?..........13
What subject matter is covered in the tests?..........14
What norms are used?..........15
What demographic information is available for population standardization?..........19
Summary..........22

Standardizing for Sex, Income, and Race..........23

Why SIR adjustment?..........23
"SIR" adjustment methods..........28
Toward data collection on outcomes..........38
Tentative findings..........39

References..........41

Appendix 1: Recommendations of an Ad Hoc Committee on Measurement of Educational Competence..........47

Appendix 2: Characteristics of Selected Statewide Testing Programs..........50

Appendix 3: Basic Testing Programs in Major School Systems..........52

Tables

Page

1. Reading Test Results, by County....27

2. Hypothetical "SIR" Characteristics of a State and 2 School Districts....30

3. Population and Grade Equivalent Data: Standard and School District 1....32

4. Unadjusted or "Crude" Grade Equivalent Levels for Reading Tests, Grade 3....33

5. SIR Differences in Arizona....34

Chart

Mathematics Achievement Scores by Grade Level Equivalent for Pupils in Grades 6, 9, and 12 by Race and by Socioeconomic Status of Parents 24

SUMMARY

The increasing use of educational performance or outcome measurements for a range of policy purposes points to new procedures for adjusting data for population composition. The purposes include:

-- program formulation

-- budget resource allocation

-- grant-in-aid designs

-- performance incentive payments

-- consumer information for school selection

-- program evaluation and review.

This paper outlines methods for controlling population differences to make data on performance more comparable across time and from place to place. The demographer's tool of population standardizations has been forged anew to meet the special problems of school performance.

The resulting estimates of achievement scores, standardized for population differences, are useful for comparison only. Such comparative indexes remove the influence on average scores of population changes over time, or population differences between schools or school districts. Adjusted scores are not intended to take the place of the basic data but to complement them.

Standardization procedures can be applied to achievement test scores and to other measurements of competence such as attitudes or attributes. In this report, achievement score adjustment is used as an example. The selection of sex, income, and race (SIR) as control variables is proposed as a first step.

A PROPOSAL FOR A "SIR" ADJUSTED INDEX OF EDUCATIONAL COMPETENCE

INTRODUCTION

Achievement test scores are now widely used as measures of educational performance and outcomes. There are, however, problems in interpreting the scores of student populations with different demographic characteristics. For some purposes, such as measures of pupil progress, individual scores may be useful. For other purposes, such as comparing the performance of educational units such as schools or school districts, they can be misleading.

Direct comparisons of the performance of different schools or school districts become, essentially, comparisons of incomparables when pupil populations differ in demographic, socioeconomic, or "cultural" characteristics. Adjustment of school population scores for differences in sex, income, race, and age--when appropriate--would reduce the bias in intertemporal and interjurisdictional comparisons.

School districts are now comparing one school's performance to another's. The scores of pupils in particular grades in a city are compared with national grade equivalent norms. At the time this report was being proposed, for example, the Chicago press in June 1971 headlined "Chicago's Pupils Get Poor Test Grades. . . . The Citywide norm of 32 falls 18 points below the national norm of 50."(1) The size and direction of the gaps between the city mean scores and national norms currently define the quality of a city's schools. State systems, moreover, are making dollars per pupil and other comparisons between school districts the traditional index of input allocations and reading or mathematics achievement scores as indicators of educational output.

In reality, because of the different student population characteristics associated with separate educational units, it is increasingly difficult to judge from test score medians whether the school or school district performance is in fact lower in one place than in another, and whether an improvement has been made from one period to another. Changes in student population composition could introduce variations in statistical results which have nothing to do with differences in the quality of education. It may be unwarranted to assume, for instance, that scores that were lower one year than the previous year necessarily reflect qualitative deterioration in the educational program. In some instances, favorable quantitative results do not really indicate a concurrent gain in quality. Real changes in results may be obscured by changes in the characteristics of the population; even though there is no change in overall test results there may actually be qualitative improvement in some instances, and decline in others.

Differences and changes in demographic characteristics of States, school districts, or schools may create statistical artifacts rather than depict real trends; achievement score data at any particular time may show interstate or school district variations that may or may not indicate comparative "learning" achieved. Even in two neighboring schools--one with only girls in attendance, the other with only boys--median or mean test scores could differ; yet each school might have scores equivalent to the national median or mean for each sex.

In the past, when children with one set of characteristics have not achieved as well as others on standardized tests, one of three courses of action has been pursued: (1) attempts have been made to reduce differences between subpopulation groups by removing bias in tests; (2) multiple rather than single tests and test norms have been developed; or (3) testing has been stopped. Early in the history of intelligence testing (IQ tests), questions were screened for differences in the response of girls and boys; if significant differences were found in responses to particular questions, items were deleted from the test. This suggests a route that might have been followed in achievement testing when any question or group of questions elicited significant differences between blacks and whites or other groups. Research findings that the origin of much difference is cultural-linguistic[2] indicate correction of tests is in order.

Through the National Assessment of Educational Progress, educational materials are being prepared that can perhaps break from the traditional white middle-class male biases in testing instruments. Exercises dealing with black culture, black history, and black literature are included. The testing exercises are objective or criteria referenced and are not normed. And material is presented so as not to compound difficulties; for example, tests not intended to measure reading comprehension are administered orally. Similarly, exercises are being examined for possible sex bias; for example in content of science questions. Such efforts to rid tests of cultural and other bias may be a step toward more accurate assessment of achievement.

A second course is to develop separate tests or use separate norms for differing groups. Separate norms for girls and boys have been used over many years but different norms for blacks and whites have come to be applied only recently and then only in connection with use of test scores for college selection. Multiple norms also have been developed. For example, norms have been set for big-city school systems to relate scores for one core city to those for another, with big-city equivalents generally 25 percent below national norms. Such norms add considerable depth to our understanding of achievements by school or school district, and provide more reliable statistical yardsticks with which to measure comparative progress. While the measurement of intra-city and intercity differences and rates of progress are improved, differences and changes in the underlying characteristics of the school population can obscure the meaning of established norms. For example, the meaning of test scores is affected by selective migration to cities, particularly for smaller school districts with special characteristics of "movers" and "stayers" in the student population.

What type of statistic would facilitate achievement score and other competence comparisons across time and across jurisdictions without adding to the display overload? The remaining sections of this paper address this question.

PURPOSES OF INTERJURISDICTIONAL AND INTERTEMPORAL COMPARISON

We ask first, what are the purposes of statistical comparisons of outcomes?

A number of differing purposes are propelling the Federal Government, States, and school districts to assess achievement in the schools. In turn, those assessments are altering structures and policies in elementary and secondary education.

Are school systems assuring equality of opportunity?

The survey of Equality of Educational Opportunity ("The Coleman Report"), carried out by the U.S. Office of Education under the Civil Rights Act of 1964, sharpened the focus on school "results."(3) Various measures of results identified in the context of racial equality include: (a) occupational status and mobility, (b) years of schooling, and (c) income.

The Coleman study was based on a sample of 564,000 children in grades 1, 3, 6, 9, and 12. The children were tested on verbal ability, nonverbal intelligence, reading comprehension, mathematics, and general material including practical arts, humanities, natural science, and social science. These intermediate educational achievements were considered necessary to ultimate occupational, educational, and income attainments. Uniform tests were given to all groups with the resultant familiar findings: the average performance of minority pupils, except the Oriental group, was significantly below the average for white students; school inputs apparently compensated little for handicaps in home and community environments.

The President in his March 1970 statement on Elementary and Secondary School Desegregation called anew for equal educational opportunity. In recommending added funds, the President said.

> I am not content simply to see this money spent, and then count the spending as a measure of accomplishment. For much too long, national "commitments" have been measured by the number of Federal dollars spent rather than by more valid measures such as the quality of imagination displayed, the amount of private energy enlisted or, even more to the point, the results achieved. (4)

What do children learn?

Another question posed by the President, in his March 1970 message on educational reform, gave new emphasis to the outcome of schooling and new measures of achievement.(4) He proposed that a National Institute of Education take the lead in developing new measurements of educational output. "NIE . . . would develop criteria measures for enabling localities to assess educational achievement and for evaluating particular educational programs. . . . In doing so, it should pay as much heed to what are called the 'immeasurables' of schooling (largely because no one has yet learned to measure them) such as responsibility, wit, and humanity as it does to verbal and mathematical achievement."

Subsidiary but related questions ask: What do children learn compared to what they could be learning? Are children learning more now than children did years ago?

How are the schools to be held accountable for their performance?

Within a surprisingly short period of time the concept of "school outcomes" has come to be applied as administrative measures of performance of schools, teachers, school districts, and so forth. Accountability has come to be a part of current practice grounded in the evaluation provision of Title I of the Elementary and Secondary Education Act (ESEA) and encouraged further by new programs such as Right to Read.

The President's March 1970 education message noted: "School administrators and school teachers alike are responsible for their performance, and it is in their interest as well as in the interests of their pupils that they be held accountable . . . . Success should be measured not by some fixed national norm, but rather by the results achieved in relation to the actual situation of the particular school and the particular set of pupils."[4] Later, in his 1974 budget message, the President set the pattern of government-wide responsibility for program performance. Programs will be evaluated to identify those that must be redirected, reduced, or eliminated because they do not justify the taxes required to pay for them. Federal programs must meet their objectives, and costs must be related to achievements.[5]

News reports on how good a job a school does are a direct consequence of evaluation in school districts. When programs are evaluated, results have to be made clear and simple. Evaluation of a program requires a clear statement of purpose and a measure or measures that quantify the essential character of that purpose. Many States have applied management by objectives to education, along with cost-benefit principles embodied in some form of planning-programming-budgeting system (PPBS), and have tied statewide educational assessment into such a system. New York is reported by ETS* to employ an adaption of PPBS--Program Analysis and Review (PAR)--to designate educational problem areas directly applicable to the State's ESEA programs. California also has been developing a PPB system.[6]

Program evaluations have heightened interest in concepts of program outputs and in data that can illuminate those concepts. Achievement testing by schools and school districts has been encouraged by evaluation requirements and some States have conducted statewide testing.

* Educational Testing Service

Achievement testing is only one of many measures that might be made of competence--both affective and cognitive--created through education. The emphasis on achievement testing, and in particular on reading scores, represents an early and undoubtedly too simple response to the need within a program evaluation to equate measure to purpose, reading score to achievement level, test result to educational outcome. (7)

Evaluation and accountability have spawned still another species--the "performance contract"--which pays contractors, teachers, or students according to student performance. (8) A whole new area of contract purchases for student learning permits industry to serve the schools by designing learning instruments, curriculum materials and the like. Payments according to performance have sharpened concern about outcome measurements; the process of evaluating performance has made it unmistakably plain how little is known about educational outcomes and about ways of achieving educational performance.

What is education's role in social accounting?

Social accounting, similar in concept to GNP accounting, has received much attention. Development of human capacity is so much a part of well-being that a measure of education is necessarily a central variable in any index for social accounting. As a step toward charting the Nation's social progress, a social report was developed in 1968 as a trial effort to "examine the qualitative condition of society regularly and comprehensively." (9) The report emphasized the need for a national assessment of educational achievement. That assessment, now underway through the Education Commission of the States and the National Center for Educational Statistics, is beginning to provide National data on the achievement of specified objectives.

Toward a Social Report (9) noted that The Digest of Educational Statistics contains over a hundred pages of educational statistics in each annual issue, yet has virtually no information on how much children have learned. The former report measured educational progress by indicators of equal opportunity such as relative positions in society and of society's enrichment by learning.

Among the measurements of equality were: (a) changes in occupational patterns, (b) years of schooling completed, (c) talent loss (percentage of persons who graduate from high school but do not go on to college), and (d) intergenerational upward mobility.

Among the measurements of enrichment were: (a) years of schooling, (b) rates of functional illiteracy, (c) school performance changes over time (using standardized test scores such as the PSAT* and SAT** scores and professional test score results), and (d) closeness between black and white achievement test scores (Coleman study).

The National Goals Research Staff's (NGRS) 1971 report to the President (10) did not undertake a second round toward a social accounting; it did note ongoing work in the Office of Management and Budget (OMB) to improve measurement of the "domestic health" of the Nation. Among the data assembled by the NGRS to open emerging issues for discussions were: (1) enrollments over time by level of education, (2) years of schooling completed by population cohort ages 35 to 39, and (3) voter responses to schools, as evidenced by public school bond election results. OMB is preparing a publication on social indicators that groups existing educational data under two social concerns: basic skills for everyone, and opportunity for advanced learning. (11)

* Preliminary Scholastic Aptitude Test

** Scholastic Aptitude Test

Lack of data rather than lack of interest in intertemporal comparisons caused recent works on social indicators to neglect comparative achievement scores.

STATUS OF ACHIEVEMENT TESTING IN STATE AND COMMUNITY

The increasing importance of knowledge about educational outcomes in policy formulation and decision-making has created a rising demand for measures that can provide that knowledge.

What kinds of information on educational achievements are now available that could be collected by NCES from schools, school districts, or States? The answers to six related questions would determine whether test score data can be collected without new surveys:

1. What portion of school children are now routinely tested?
2. How many different tests are given?
3. At what grade levels are the achievement tests given?
4. What subject matter is covered in the tests?
5. What norms are used?
6. What demographic information is available for standardization?

Information addressing these questions is drawn from several sources: Surveys made by the Educational Testing Service in 1968[12] and 1971;[6] the Akron Public School Survey in April 1968 of basic testing programs used in major school systems throughout the United States;[13] and the 1970 survey of the Research Council on Greater City Schools.[14]

The ETS studies are specifically concerned with State testing programs, defined as any organized, coordinated, centralized effort by a State to provide some type of test materials or services. The definition, however, includes States furnishing every conceivable service associated with testing and States that merely offer assistance in developing or improving local testing programs.

What portion of school children are now routinely tested?

Educational testing in the States has been encouraged by Federal requirements for evaluation. The growth has been accelerating. Informational materials for the ETS 1967 study were submitted by 50 State departments of education and a selected group of colleges and universities. In responses for that year 42 State departments reported testing programs; eight states indicated no programs. Most of the programs were intended principally for guidance of students. Only 17 States were using tests to help evaluate instruction and only 13 to assess student progress.

1. At least 2 million pupils in 10 states are tested annually by at least one of these five tests: California Achievement Test, Stanford Achievement Test, Iowa Test for Basic Skills, Metropolitan Achievement Test, and Science Research Associates Achievement Tests.

2. Almost 6 million additional pupils are tested in extensive State testing programs in other States.

The recent requirement for program evaluation under Title I of the Elementary and Secondary Education Act has greatly increased the use of achievement tests in States and communities (with in some instances separate reporting by sex, family income, and race of pupils). The newer Right to Read program has also stimulated achievement measurement. There has been increasing concern over the kinds of measurable pupil learning and development which State educational tax dollars are buying. According to the 1971 ETS compilation of State Educational Assessment Programs,(6) every State had conducted a needs assessment program, was currently doing so, or planned to recycle a completed one. The universal use of such programs, ETS felt, was explained by the requirement of section 402 of ESEA, Title III, which tied needs assessment to the receipt of Federal funds.

In addition to the individual State programs, 27 States had participated in planning the Belmont System, a comprehensive educational evaluation system developed with the cooperation of the U.S. Office of Education to help consolidate and improve State reporting required by law under several Federal aid programs.

Many States are emphasizing the formulation of statewide educational goals, in recognition that such a set of goals is an essential characteristic, if not prerequisite, of an educational assessment program.

How many different tests are given?

The ETS survey on State testing programs reported achievement testing batteries in 27 States in 1967 as follows:

	Programs	States	Different Instruments
Achievement batteries	34	27	21

The ETS survey gave the type of tests and the number of States in which each testing instrument is used. Most of the children tested annually are tested under local programs. The following figures were computed from the 1971 ETS survey of State educational programs:

Tests	Programs	States
Iowa Tests of Educational Development (ITED)	11	11
Stanford Achievement Tests (STAT)	9	8
Sequential Tests of Education Progress (STEP)	5	5
California Achievement Tests (CAT)	5	5
Iowa Tests of Basic Skills (ITBS)	7	6
Metropolitan Achievement Tests (MAT)	2	2
Science Research Associates Achievement Series (SRA)	5	5
	44	42

At what grade levels are the achievement tests given?

Achievement testing is usually done by grade level. Mental Measurements Yearbook[15] reports possible ranges:

	Grades
Iowa Tests of Educational Development (ITED)	9-12*
Stanford Achievement Tests (STAT)	1.5-2.4, 2.5-3.9, 4.0-5.4, 5.5-6.9, 7.0-9.0
Sequential Tests of Education Progress (STEP)	4-6* 7-9 10-12 13-14
California Achievement Tests (CAT)	1.5-2.0 2-4 4-6 6-9 9-12
Iowa Tests of Basic Skills (ITBS)	3, 4, 5,* 6, 7, 8-9

*Only in 1965 volume of Mental Measurement Yearbook

	Grades
Metropolitan Achievement Tests (MAT)	K-1.4 1.5-2.4 2.5-3.4 3.5-4.9 5.0-6.9 7.0-9.5
Science Research Associates (SRA)	1-2 2-4 3-4 4-9

Some States, notably Michigan and Pennsylvania, have set in motion programs of statewide testing in several subject matter areas and others, such as Colorado and Delaware, are moving in that direction. Some States are starting unit testing for a grade level and others restrict the tests to reading.

As of 1971, for grade groupings, the following numbers of States reported testing programs:

Grade levels tested	Number of States testing
K-3	13
4-6	24
7-9	22
10-12	22

Special characteristics of State testing programs are summarized in Appendix 2. Batteries of tests are often given in selected large cities. A 1970 survey by the Research Council on Greater City Schools reported testing in 100 major city school systems. Findings of an Akron study on tests by grade level are shown in Appendix 3.

What subject matter is covered in the tests?

The subject matter varies by grade and test. Word meanings, vocabulary, reading comprehension, and arithmetic computation are among the subjects most often included.

Although the central purpose in most States is to assess the cognitive development of students, a few States are beginning to stress personal-social development as well. Pennsylvania includes attitudes and noncognitive abilities that it has set as part of the schools' purposes: among the tests are measures of self-concept, understanding of others, citizenship, creativity, health habits, readiness for change, and attitudes toward the school. Michigan has measured attitudes toward learning, achieving, and self.

What norms are used?

Various types of norms developed by test publishers are being applied in comparisons of schools, school districts, and so forth. nationally standardized tests establish norms from responses to the tested material by a national sample of the school population.

Norms for tests are developed on the basis of either raw scores--counts of correct answers--or derived scores--sets of values which describe the test performance by some specified group, usually shown as a table giving equivalent values of some derived score for each raw score on the test, with the purpose of:

1. making scores from different tests comparable by expressing them on the same scale, and/or

2. making possible more meaningful interpretations of scores.

Types of derived scores for interindividual comparison in standardized achievement tests include:

A. Transformations based on the mean and standard deviation of the scores for the group (linear standard scores):

(1) Z-scores*
(2) AGCT-type scores (Army General Classification Test)
(3) CEEB scores (College Entrance Examination Board)

B. Transformations based on relative position within group:

(1) Rank
(2) Percentile ranks and percentile bands
(3) Stanines
(4) T-scores**
(5) Normalized standard scores

C. Consideration of the proportion of possible test scores:

(1) Percent placement

D. Consideration of the status of those obtaining same score:

(1) Age scores
(2) Grade-placement scores

For most tests, publishers provide national norms; for some, regional norms are available; and in those States in which State testing has been done for some time--New York, Alabama, California, Iowa, Rhode Island, Minnesota, Pennsylvania, Michigan--statewide norms are available.

<u>Problems of variations and biases in norms</u>. Roger Lennon's discussion of norms in a 1963 ETS paper notes:

> There are good reasons for supposing that differences in norms ascribable simply to . . . variations in norming procedures are not negligible. When we consider that to such differences from test to test there must be added differences

* Z-score or transformed standard score is a modified standard score developed to avoid decimals and negatives.

** T-score or normalized standard score is a score that would be equivalent to the score if the distribution had been normal.

> associated with varying content, and with the time at which standardization programs are conducted (including the time of the school year), the issue of comparability, or lack of it, among the results of the various tests may begin to be seen in proper perspective. Empirical data reveal that there may be variations of as much as a year and a half in grade equivalent among the results yielded by various achievement tests; variations of as much as 8 to 10 points of IQ among various intelligence tests are, of course, by no means uncommon. (16)

Existing norms, in addition to being often outdated, suffer from samples which may differ systematically from the National population.

In data gathering, norm biases are particularly important since they distort single-time and longitudinal comparisons among schools, school districts, and States. The reduction of norm bias is critical because these are the kinds of comparisons most often sought.

Translation of different test scores. NCES has completed an ANCHOR Test study to develop score correspondence among the seven most used reading tests (with an eighth being developed). Score correspondence is essential to any nationwide data collection effort that leaves to the local community and State the initial decision on what children should learn and are learning--and the testing instruments to assess that learning.

A feasibility survey was launched in 1969 on reading comprehension subtests for the five most widely used standardized test batteries, appropriate for grades four, five, and six. The reading comprehension subtests of the Metropolitan Achievement Test, Stanford Achievement Test, Iowa Test of Basic Skills, SRA Achievement Series, and the Sequential Tests of Educational Progress were administered to some 830 children, each completing subtests from three batteries arranged in random order. Correlation coefficients among the five subtests were high: the lowest (for groups of grade four pupils) was 0.81 and the highest (for the same grade) was 0.91.

In another feasibility study mathematics test scores could not be translated from one test to another.

Based on the results of the reading feasibility survey, a major test-equating and standardization study was conducted. The number of tests for which correspondence was sought was expanded among test instruments. The purposes of the study were:

1. to set up nationally representative norms for reading comprehension, vocabulary, and total reading scores for the most widely used form of the 1970 version of the Metropolitan Achievement Test, at levels appropriate for grades 4, 5, and 6;

2. to develop tables of score correspondence between the reading portion of the Metropolitan and corresponding subtests of 6 other test batteries;

3. to formulate new, nationally representative norms for the reading comprehension and vocabulary subtests of the other reading tests (5 or 6); and

4. to estimate parallel-form reliabilities for reading comprehension and vocabulary subtests of the test patterns.

Criterion-referenced testing. A move away from norm setting has come to be urged, partly because of minority group reaction, partly because of the general overall policy uses of testing results, and partly as a consequence of the increasing acceptance of the National Assessment of Educational Progress. Criterion- or objective-referenced tests are designed to measure performance on clearly stated objectives that identify specified skills in a particular subject matter area. Scores on such tests show the percent of correct responses on each of the pieces of information identified "as important to know" in accord with the identified objectives. Despite the growing acceptance of criterion-referenced instruments, problems remain of summarizing results of percent correct responses when tests in different subject matters have a varied range of questions (easy to hard) and when questions within one testing instrument are of unequal difficulty. (17)

What demographic information is available for population standardization?

The development of comparative indexes of achievement requires that information be available on demographic characteristics, at least sex, income or parental education, race, and possibly age for each pupil for whom achievement scores are available. Data are required on those variables that approximate the direct SIR (sex, income, and race) data for the school unit population of children tested.

The availability of information on achievement tests and on SIR suggests that for the particular achievement tests given: (1) sex and age data are probably available for almost all States and communities; (2) race is somewhat less likely to be available; and (3) income data (of varying qualities) are available in some States and communities but not in most others.

Data on income cannot be obtained directly or easily from students or teachers, and inquiries to parents may be misinterpreted as prying by school officials. Occupational category or property values in the school neighborhood might be used as indirect income indicators, as could educational or occupational status of parents.

Internal Revenue statistics on income for a small area could be related to Census tract data and, in turn, to school district data. Recently, in connection with Revenue Sharing administration, tax forms were amended to permit routine income data tabulation; and data matching pilot projects are needed to assure convertibility from Census to Internal Revenue sources. And consideration might be given to such indirect measurements as an area-wide socioeconomic status score, based on occupation, education, and income, which was developed in 1960 by the Census Bureau.

Census tract data from the 1969-70 Census of Population are available on race, sex, and family income. The NCES-sponsored project for mapping Census tracts with ELSEGIS* districts facilitates the matching of school district-data on achievement with a wide range of Census socioeconomic information; however, obtaining similar data within school districts poses a major problem.

For selected time periods, a national analysis is practicable for achievement scores and changes showing separately achievement data by sex, by income, by race, and by other characteristics. Sources of such national data include some 20 surveys now being compiled for NCES. However, the data are too sparse to warrant adjustment for population subgroups. Only as added information becomes available over the years through routine fact gathering are summary statistical methods such as SIR Adjusted Index indicated.

National Assessment materials becoming available show differences in achievement scores by sex, race, and economic status

* Elementary and Secondary Education General Information Survey.

(as measured by the highest educational level achieved by either parent). The purpose of the assessment is to provide information about progress in the achievement of educational objectives. In designing this program. objectives and corresponding test exercises were carefully reviewed by scholars, educators, and lay citizens. The testing instruments already completed or scheduled provide samples of exercises appropriate for four age groups (9, 13, 17, and 26-35) in 10 subjects: science, reading, writing, citizenship, art, career and occupational development, literature, mathematics, music, and social studies.

The exercises measure knowledge, skills, and attitudes of groups, rather than individuals, according to their: (1) age level; (2) size and type of community (extreme inner city, inner city fringe, extreme affluent suburb, suburban fringe, medium city, extreme rural, and small cities); (3) four geographic regions of U.S.; (4) socio-educational levels; (5` race; and (6) sex. Thus, the assessment data provide the framework for required adjustments over time and indexes based on a standard population.

National Assessment exercises differ from the standardized tests in that the goal is estimates of group rather than individual performance. No individual answers all the questions in each testing instrument; different groups of questions are administered to different samples of the population, as in public opinion polls. The intent of the summary statistics draw from the assessment is to indicate what percentage of the population or of population subgroups can answer specific questions according to the predetermined criteria.

Additional sources of data include:

(a) Project TALENT, carried out initially in 1960 as a National sample survey of half a million high school students (grades 9 to 12), measured human talents with a series of specially constructed and tested measurement instruments. Information was obtained on such student characteristics as income and sex. Race was only reported as a school characteristic so that data on achievement scores and race cannot be tied directly.(18)

(b) The Health Examination Survey of 1963-65 included in its second cycle a 60-minute test battery to assess mental aspects of growth and development of 6 to 11 year olds. The Reading and Arithmetic subtests of the Wide Range Achievement Test were given to measure school achievement. Findings in their raw score form have been presented by age, grade, and sex. Grade equivalents, percentile ranks, and standard score equivalents of the raw scores are also presented.(19)

(c) NCES study on reading test measurements in grades four, five, and six will provide, in addition to test score correspondence, data on race and sex of respondents and their family income (as judged by the classroom teacher).

Summary.

Our review of existing school testing--and data for adjusting statistical summaries for sex, income, and race--show that:

(1) In some States or communities, achievement score data are available for selected grade levels and a beginning is being made on measurements of other competencies. In a smaller number of States, data on income, sex, and race are also available.

(2) For national cross-time comparisons, National Assessment data are becoming available with data on sex, race, and economic status (as measured by parental education).

(3) Complete State-by-State and school district data are not now available, nor are the existing data comparable because of the variety of tests given and the range of grades at which tests are administered.

STANDARDIZING FOR SEX, INCOME, AND RACE

Why SIR Adjustment?

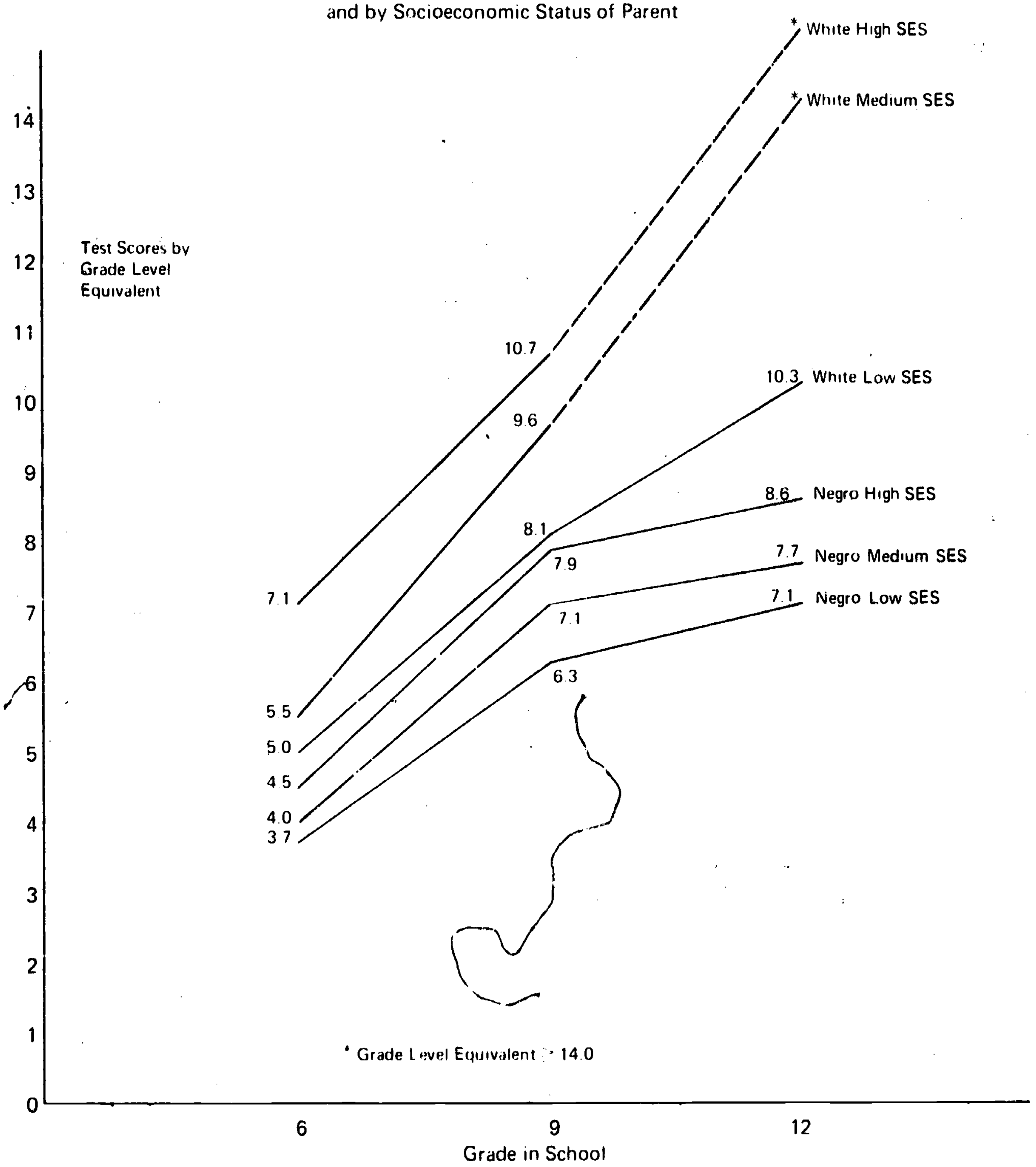
The differences in test scores by socioeconomic status and race are easy enough to display when the amount of information is limited. The chart on page 24 shows mathematics test scores in grade level equivalents for pupils in the 6th, 9th, and 12th grades and the movements of test scores for whites and blacks in high, medium, and low socioeconomic status (of parents).

It would be difficult indeed to show on the same chart or table test scores in grade level equivalents by race and socioeconomic status of parents over time. Because of information overload the information must be summarized so that differences can be seen more readily. Standardization for population is a familiar way of displaying information comparably across jurisdictions or across time.

Standardization techniques are well known, primarily in demographic studies concerned largely with birth and death rates, but also in other areas such as labor force participation. Similar statistical methods could and should be applied to achievement scores and other measures of educational outcomes on competence.

Chart

Mathematics Achievement Scores by Grade Level Equivalent for Pupils in Grades 6, 9, and 12 by Race and by Socioeconomic Status of Parent



Source: Based on data in George W. Mayeske, et al., "Growth in Achievement for Different Racial, Regional, and Socio Economic Grouping of Students," U.S. Office of Education, May 1969 (processed).

A single set of national norms for all girls and boys, all income groups, and all racial groups would be appropriate only if one assumes that all groups: (1) have the same interests, (2) are exposed to the same learning experience, (3) have the same opportunity to learn, and (4) have the same verbal competencies.

It is frequently maintained that an atypical pupil or an atypical group of pupils (atypical in terms of educational opportunities) cannot be "fairly" judged by a test which assumes equal educational backgrounds. Many standardized tests, for example, do not differentiate norms by sex; girls, however, tend to score higher on tests that are verbal in nature while boys tend to score higher on tests that are numerical or mechanical.

Suppose we say that reading for understanding with some competence is a basic skill that all children should master. There still remains a set of facts that would lead us to correct scores of achievement tests for the special characteristics of the school's population. The vocabulary of one group may differ from that of another; the same word in fact may convey very different meanings-- a "fly," a "strike," a "cat." If school districts or schools are being compared--in one, children are from homes with highly developed formal English language skills, in the other, formal English is not a first language and there is little family participation in child education--comparisons would hardly be useful for assessing progress in student language skill achievements. Even if test instruments were free of verbal biases, it might be desirable to have separate reference norms for girls and for boys, for blacks and for whites, for rich and for poor.

We are not disputing the argument that minimal requirements of basis skills should be applicable to all. As the President indicated in his 1970 message to Congress:

> For years the fear of "national standards" has been one of the bugaboos of education. There has never been any serious effort to impose national standards on educational programs, and if we act wisely in this generation we can be reasonably confident that no such effort will arise in future generations. The problem is that in opposing some mythical threat of "national standards" what we have too often been doing is avoiding accountability for our own local performance. We have, as a nation, too long avoided thinking of the <u>productivity</u> of schools
>
> This is a mistake because it undermines the principle of local control of education. Ironic though it is, the avoidance of accountability is the single most serious threat to a continued, and even more pluralistic educational system. Unless the local community can obtain dependable measures of just how well its school system is performing for its children, the demand for national standards will become even greater and in the end almost certainly will prevail. When local officials do not respond to a real local need, the search begins for a level of officialdom that will do so, and all too often in the past this search has ended in Washington.(4)

Study is under way on the problems of comparing school districts and States on the basis of test performance. Test score interpretation based on differential norms is useful in comparing school districts with comparable characteristics. But other methods are necessary when comparing systems with different characteristics.

Henry S. Dyer has proposed a method of computing a School Effectiveness Index (SEI) that "automatically adjusts for the differing circumstances in which a school must operate." His educational accounting system has a procedure for establishing SEI profiles for a school. The procedure calls for a series of regression analyses, using the test scores and background characteristics from all schools in the area within which comparisons are to be made. Measures of quality are then determined by the distance of a school from the regression line.(20) Dyer identifies hard-to-change as contrasted with easy-to-change variables (or circumstances in which schools must operate). By controlling statistically variables over which schools have little or no control, he points out, a school and its staff are better able to determine how effective their efforts may be. SEI's may indicate ways in which a school staff might improve its performance.

To illustrate the problem furtner, we drew on 1970-71 data from Arizona showing reading test results by county (table 1) in a test given to third graders.(21) Average grade equivalents are below the third-grade level in all counties except Yavapai, with a range of 2.6 to 3.0. Assuming there is a significant difference between 2.6 and 3.0, is it a result of differences in educational quality or is it the result of differences in population characteristics in Yavapai on the one hand and Apache on the other?

Table 1. Reading Test Results, by County

County	Average Grade Equivalent
Apache	2.6
Cochise	2.9
Coconino	2.8
Gila	2.9
Graham	2.8
Greenlee	2.9
Maricopa	2.9
Mohave	2.9
Navajo	2.8
Pima	2.8
Pinal	2.7
Santa Cruz	2.6
Yavapai	3.0
Yuma	2.8

"SIR" Adjustment Methods

In comparing any two school districts, States, other population groups, or the same population group or community at various points of time, control for differences in race, sex, and income distribution permits a more realistic picture of educational outcomes. Standardized outcome indexes are meaningful only for comparison. But the adjusted scores, when used in conjunction with unadjusted ones, contribute greatly to an understanding of the data for comparing educational achievements in different places, or at different times.

The most common methodology in population standardization is to assume a standard set of demographic characteristics for all areas, or all dates being studied. Specific rates or achievement scores for each of the subpopulations in each area or at each time period are then applied to the standard population. This calculation would show, for example, achievement scores which would have been experienced if the sex, income, and race characteristics of the school district had been the same as those in a standard population. Since the standard population is applied to all communities or time periods being studied, differences in race, sex, and income composition are removed or are held constant in making the comparisons. The specific mechanics for computing standardized rates can vary. Two general methods are outlined here: <u>Community average specific achievement scores weighted by a standard population</u> and <u>An index of educational achievement in which the specific achievement scores for a standard population are compared with the standard scores for a standard population</u>.

Community average specific achievement scores weighted by a standard population. For example, assume a standard population characteristic in a State with the following racial composition: white, 55 percent; black, 16 percent: other (including American Indian), 29 percent. The sex distribution is somewhat weighted in favor of females at 51 percent. Income is shown in three classes only (a classification that seems too undifferentiated, but represents existing practices within a State)., The income classes show 30 percent of the population with incomes under $3,000; 60 percent with incomes between $3,000 and $10,000; and 10 percent with incomes $10,000 and over.

Table 2 (see page 30) illustrates the components of a standard population by sex, income, and race. For this particular illustration, three classes of race are indicated and three income groups.

"SIR" adjusted achievement test scores in school districts in the State can be calculated by multiplying average specific scores for each of the 18 subpopulation groups or categories (composed of three race categories, the two sexes, and three income classes) by the corresponding standardized population distribution for each subpopulation. In School District 1 the white population is shown as a larger percentage of the total than in the standard population, and the black a substantially smaller percentage. By the same token, a smaller proportion of the population is in the under-$3,000 income class with an average score of 2.6, and more than double the standard population is in the $10,000-and-over income class with an average score of 5.1. School District 2's characteristics are assumed to come closer to those of the standard for the State as a whole.

Table 2. Hypothetical "SIR" Characteristics of a State and 2 School Districts

(In percent)

Standard Population Characteristics in State

White	- 55	Male	- 49	Under $3,000	- 30
Black	- 16	Female	- 51	$3,000-$10,000	- 60
Other	- 29			$10,000 +	- 10
	100		100		100

Characteristics of School District 1

White	- 80	Male	- 49	Under $3,000	- 17
Black	- 08	Female	- 51	$3,000-$10,000	- 62
Other	- 12			$10,000 +	- 21
	100		100		100

Characteristics of School District 2

White	- 60	Male	- 50	Under $3,000	- 35
Black	- 10	Female	- 50	$3,000-$10,000	- 55
Other	- 30			$10,000 +	- 10
	100		100		100

There is no right number of population characteristics to be used in standardization. The combinations of specific rates depends basically on the number of groupings or classifications considered useful for adjusting for sex, income, and race. For example, average grade equivalent information for Arizona differentiated five racial groups, showing average grade equivalents for third-year reading tests for each group. In other States the Indian population, for example, or those with Spanish surnames, may not account for so large a portion of the population as to warrant separate classification.

Income is estimated in the State of Arizona by teachers. The data are compiled in the State on the basis of teacher information for three broad income groupings: smaller spreads and thus more classes of income might be desirable in showing achievement score differences for standardized populations. The difficulties of greater precision in income reporting relying on teacher observations are great if not insurmountable.

In all, some 30 combinations of SIR-specific scores might be a reasonable number of combinations that would permit achievement scores to be shown as SIR-specific rates and for which averages could be computed and reweighted for comparative purposes. This number would consist of three racial subgroups, five income classes and two sex groups.

There is no right way of selecting a comparative standard for demographic groups across either time or communities. The standard depends primarily on the units subject to comparison. If, for example, school districts within a State are compared, the standard could be the average race, income, sex composition for the State as

Table 3. Population and Grade Equivalent Data: Standard and School District 1

	White		Black		Other	
Income	% of total pop.	Grade equiv. levels of 3rd-year reading test	% of total pop.	Grade equiv. levels of 3rd-year reading test	% of total pop.	Grade equiv. levels of 3rd-year reading test
			STANDARD			
			Female			
Under $3,000	3	2.9	4	2.0	8	1.9
$3,000-$10,000	22	3.5	3	2.7	6	2.6
$10,000 +	3	4.3	1	4.0	1	3.6
			Male			
Under $3,000	3	2.4	4	1.8	8	1.7
$3,000-$10,000	21	3.2	3	2.6	5	2.5
$10,000 +	3	4.1	1	3.8	1	3.6
			SCHOOL DISTRICT 1			
			Female			
Under $3,000	4	3.6	2	2.5	3	2.3
$3,000-$10,000	29	4.4	1	3.4	2	3.3
$10,000 +	8	5.4	1	5.0	1	4.5
			Male			
Under $3,000	3	3.0	2	2.2	3	2.1
$3,000-$10,000	27	4.0	1	2.6	2	3.1
$10,000 +	9	5.1	1	3.8	1	4.5

For illustrative purposes the comparison is of School District 1 with the statewide average as a standard.

a whole (which is the process assumed in the examples shown), or any one of the school districts could be used as a standard against which other school districts would be contrasted, or a national standard could be applied. Standardization of demographic characteristics for intertemporal comparisons permits the use of the standard population in a base year, the latest year, or some intervening year.

Grade equivalent scores for each of the 18 subpopulation groups of the standard and District 1 populations are shown in table 3 (page 32). In this table of hypothetical third-grade reading scores, families with incomes of $10,000 and over in the standard population have grade scores averaging 4.3 for the whites, 4.0 for the blacks, and 3.6 for other races. For the under-$3,000 income group in the standard population, the scores averaged 2.9 for whites, 2.0 for blacks, and 1.9 for other races. In the standard used here--the statewide average--the scores appear lower than those in School District 1.

The unadjusted averages or raw scores for reading tests at grade-three-equivalent levels are summarized in table 4 for the two school districts. Thus, the statewide standard score is shown to average 2.6; the average for School District 1, 4.1; and for School District 2, 2.9.

Table 4. Unadjusted or "Crude" Grade Equivalent Levels for Reading Tests, Grade 3

Statewide Standard Score Average	2.6
School District 1 Average	4.1
School District 2 Average	2.9

For each of the components of SIR, the kinds of differences drawing on the Arizona county data are illustrated in table 5 (page 34).

Table 5. SIR Differences in Arizona

Average grade equivalents differ statewide as follows:

By race:	
White	3.7
Spanish surnamed	2.8
Black	2.7
Indian	2.6
Oriental	4.1
By sex:	
Male	3.3
Female	3.6
By income:	
Below $3,000	2.6
$3,000 - $10,000	3.4
Above $10,000	4.2

A hypothetical school district, District 1, which on raw scores averages 4.1 for reading tests at grade three levels, has a reduced score of 3.6 when corrected for population differences; District 1 has a larger proportion of whites and/or higher income groups than is "standard" for the State. It has a higher unadjusted score than would in fact be attributed to it if it had a standard population distribution.

Standardizing for population differences thus changes the unadjusted or raw score average for School District 1 from 4.1 to 3.6. If the population distribution by sex, income, and race in School District 1 had been the same as the statewide average (if the proportion of the high income class were lower and the proportion of whites were somewhat lower), the average would be lowered from the raw average shown.

An index of educational achievement in which the specific achievement scores weighted for a standard population are compared with the standard achievement scores for a standard population. This index asks the question: Is the achievement score in the District (or State) higher than the "average" or not, and by what percentage? For time-period data the index would ask: Is the achievement score higher or lower in one year than some base period?

This measure, standardized for population, would call for the computation of an average adjusted score. This score would be the sum of specific scores weighted by the population standardized for sex, income, and race by grade level divided by the standard scores for the standard population (multiplied by 100).

Adjusted achievement scores by color, sex, and income are computed by using, in the numerator, average scores specific for color, sex, and income multiplied in each instance by the distribution of the standardized population for those characteristics.

$$\text{Adjusted score} = \frac{\text{Sum of the products of the SIR specific rates multiplied by standard population distribution}}{\text{Sum of the product of the norm or standard scores times standard population distribution}} \times 100$$

When average crude ores for a State or school district are available but scores for each subgroup are not, some indirect methods of adjustment become necessary. The two variants of the earlier methods are presented here

Variant method 1--<u>Scores adjusted for standard populations when specific rates are not available for each time period or school district for which comparisons are made</u>. Average crude scores could be adjusted on the basis of the ratio of the standard scores for a standard population to the standard scores weighted by the specific characteristics of the school population. What needs to be known is the population characteristic of the State or school district but not necessarily the specific scores for each identified subpopulation.

The crude average score in this process is multiplied by a ratio. The numerator in that ratio represents the sex, income, and race standard rates weighted by the composition of the standard population distribution (in other words, the weighted average score for the standard). The denominator of the ratio represents the average scores obtained at standard scores for sex, income, and racial groups weighted by the subpopulation distribution of the particular State (school district or time period).

Variant method 2--<u>Index of specific achievement scores compared to the standard or averages for the State or school district</u>. The variant in this instance, as in Variant method 1, is useful when average crude scores for a State or school district are available but scores for each subgroup are not.

The crude average score for a school district would be divided by standard scores weighted by population characteristic of the school district and multiplied by 100 to derive the index. The computation essentially shows the crude achievement score as a ratio of crude score to the average standard scores for the same specific population.

The two general methods of standardizing for sex, income, and race differences are summarized below, along with the variants that can be used when these methods of standardizing are used to yield specific sets of numbers:

1. Average of specific scores weighted by Standard Population.

2. $$\frac{\text{Specific Scores weighted by Standard Population Distribution}}{\text{Standard Scores weighted by Standard Population Distribution}} \times 100$$

Variant 1.

$$\text{Average crude scores} \times \frac{\text{Standard Score x Standard Population}}{\text{Standard Scores x School District Population}}$$

Variant 2.

$$\frac{\begin{array}{c}\text{Specific Scores weighted by School District Population}\\ \text{(Average on Crude Score)}\end{array}}{\text{Standard Scores weighted by School District Population}} \times 100$$

The formulas result in these numbers for the School District 1 example.

Method 1 yields a SIR adjusted rate of 3.6

Method 2 yields an adjusted score index of 138 or $(\frac{3.6}{2.6} \times 100)$.

The variant of method 1 yields 3.6 or $4.1 \times \frac{2.6}{3.0}$.

The variant of method 2 yields 137 or $\frac{4.1}{3.0} \times 100$.

Stated differently, the several indexes for School District 1 might be summarized in this way for the third grade reading score:

> . . . If there were a standard population and it had standard scores, the average would have been 2.6.
>
> . . . If school district 1 population characteristics are taken into account, but scores are standard, the average score becomes 3.0.
>
> . . . If the school district had a standard population and scores equal to its own experience, the average becomes 3.6.
>
> . . . If the school district is assessed in terms of its own scores and its own school district population, the average is 4.1.

Standardization processes can be varied further, depending upon the kinds of comparisons sought. And the comparison can be made in terms of index numbers to emphasize the comparative nature of the estimate.

For certain analyses, it is clearly desirable to apply more intricate forms of standardization; procedures outlined above involve only the application of a standard set of demographic variables to various achievement rates. Such standardization, therefore, is basically a form of weighting, and the standardized rate is a weighted arithmetic mean.

More than averages are probably needed to understand variations in educational achievement within subgroups, and it might be desirable at some later date to consider more complex adjustments.

Toward Data Collection on Outcomes

A major NCES role in educational outcomes data requires preparation for the collection of data on achieveme scores and socio-economic status of children. If educational outcomes are to be linked to program inputs and program financing, then more complete nation-wide data must be obtained. The machinery for such collection has to be built and, once designed, a strategy for implementation put into practice.

The Elementary and Secondary Education General Information Survey (ELSEGIS), conducted by NCES, after review and revision, may ultimately be the instrument for collection of data on achievement score outcomes. ELSEGIS includes a survey of expenditures and revenues of LEA's by source and account; a nationally representative sample of districts is surveyed. Earlier the Belmont Survey* of the

* In 1968 the Council of Chief State School Officers and the U. S. Office of Education undertook to jointly develop and implement a comprehensive educational evaluation system. The initial meetings took place at Belmont House in Elkridge, Maryland, and the program has been known as the Belmont project. More recently the Committee on Evaluation and Information Systems of the Council of Chief State School Officers has, in accord with one of its purposes, begun the formulation of recommendations on information required for evaluation as part of a State-local information system. (22)

Bureau of Elementary and Secondary Education undertook to collect data on finances and program evaluation materials. Some collection processes should be developed so that comprehensive data can become available for policy formulation not only at the national level but in the States and communities as well.

A March 1970 report of the Committee on Educational Finance Statistics to the U. S. Commissioner of Education paid special attention to the need for comprehensive comparative data for policy decisions.(23) While noting that the Committee had not given much attention to student achievements and attainment of other program objectives, the report noted the need for relating expenditures to educational impact and proposed data collection on educational impact as follows:

1. Number of pupils below "minimum achievement standard" (such as the fourth stanine) in reading and math per average daily attendance at various grade levels; e.g., 3. 6. and 9. (Where statewide achievement testing results are available, as in New York, Alabama, Rhode Island, Michigan, Minnesota, Pennsylvania, and California.)

2. Number of pupils below "minimum achievement standard" in reading and math per number of title I eligibles at various grade levels. (The Committee's recommendations in this area--educational impact--only illustrate types of data that, if available, would be desirable.)

For each of these common denominators, comparisons should be made between data from local schools, school districts, and State data on the same item.

The Committee, however, did not take account of the need for adjusting test scores for population differences.

Tentative Findings

1. At present a body of readily available data on achievement scores and "SIR" is not sufficient to yield State-by-State estimates

or to relate outcomes and inputs by State; selected school district data for large districts are more nearly available.

2. Extensive achievement testing is going on in the Nation's largest cities. One of six national standardized tests is being used in the lower grades of those cities.

3. None of the present methods of achievement testing and standardizing norms provides the data necessary to compare performance levels of schools or school systems with different demographic characteristics.

4. The Anchor test work provides the first mechanism for translating from one test score to another among the major reading comprehension tests.

5. Various methods need to be developed to assure that SIR data and, in particular, appropriate income data become available to match achievement testing. The Internal Revenue data appear potentially the most useful body of income information.

6. Achievement clearly is only one among several major educational outcomes. Other measures of competence should be actively pursued. The measures discussed in PSL's* report on educational outcomes demonstrate that achievement in designated subject matter is but a part of the develonment of intellectual competence.(24)

* Public Services Laboratory of Georgetown University

REFERENCES

1. "Chicago's Pupils Get Poor Test Scores." Chicago Today, June 16, 1971.

2. First National Conference on Testing in Education and Employment. Oral Discussion, Hampton Institute. Hampton, Virginia. April 1-3, 1973.

3. Coleman, James S., et al. Equality of Educational Opportunity. Washington, D.C. : Government Printing Office, 1966.

4. Messages of the President to the Congress. March 3, 1970. Message on Education Reform, March 3, 1970.

5. Budget Message of the President. The Budget of the United States Government Fiscal Year 1974. Washington, D.C.: Government Printing Office, 1973.

6. Dyer, Henry S., and Rosenthal, Elsa. An Overview of the Survey Findings in State Educational Assessment Programs. Princeton, N. J.: Educational Testing Service, 1971.

7. Levine, Robert A. The Poor Ye Need Not Have With Thee: Lessons from the War on Poverty. Cambridge: The MIT Press, 1970, p. 144.

8. Levine, Donald M., ed. Performance Contracting in Education - An Appraisal. Englewood Cliffs, New Jersey: Educational Technology Publications, 1972.

9. U. S. Department of Health, Education, and Welfare. Toward a Social Report. Washington, D.C.: Government Printing Office, 1969. p. 66.

10. National Goals Research Staff. Toward Balanced Growth: Quantity and Quality. Washington, D.C.: Government Printing Office, 1970.

11. Tunstall, Daniel. "Working Outlines for OMB Social Indicators Publication," May 1973. (Unpublished manuscript)

12. Educational Testing Service. State Testing Programs: A Survey of Functions, Tests, Materials, and Services. Princeton, N.J.: Educational Testing Service, Evaluation and Advisory Service, March 1968.

13. Akron Public Schools. "Basic Testing Programs Used in Major School Systems Throughout the United States." Akron, Ohio: Akron Public Schools, April 1968.

14. Unpublished, informal survey by the Research Council on Greater City Schools, 1970.

15. Burns, Oscar Krisen, ed. The 7th Mental Measurements Yearbook, Highland Park, N.J.: Gryphon Press, 1972.

16. Quoted by Richard M. Jaeger in "A National Test Equating Study in Reading" (processed, undated).

17. Mushkin, S. J. "National Assessment and Social Indicators." Monograph, Washington, D.C.: Government Printing Office, forthcoming.

18. Project TALENT. "A National Inventory of Aptitudes and Abilities." Bulletin No. 1, November 1959. University of Pittsburgh, Project Talent Office, Washington, D.C.

19. National Center for Health Statistics. School Achievement of Children 6-11 Years as Measured by the Reading and Arithmetic Subtests of the Wide Range Achievement Test. Vital Health Statistics. PHS Pub. No. 1000 - Series 11 - No. 103. Washington, D.C.: Government Printing Office, June 1970.

20. Dyer, Henry S. "Toward Objective Criteria of Professional Accountability in the Schools of New York City." *Phi Delta Kappa*, Vol. 52, No. 4, Dec. 1970, pp. 206-211.

21. Arizona Department of Education. *1970-71 Third Grade Reading Achievement Test Results Report*. Phoenix: Arizona Department of Education, June 1971.

22. Committee on Evaluation and Information Systems. "Bylaws of the Committee on Evaluation and Information Systems (CEIS) of the Council of Chief State School Officers." Washington, D.C., September 15, 1972.

23. Kelly, James A. "Report of the Committee for Educational Finance Statistics: Recommendations for Data Collection, Analysis and Publication." New York: Columbia University, Teachers College, March 1970.

24. Public Services Laboratory. "Educational Outcomes: An Exploratory Review of Concepts and Their Policy Application." Washington, D.C.: Public Services Laboratory, April 1972.

APPENDIXES

APPENDIX 1: RECOMMENDATIONS OF AN AD HOC COMMITTEE ON MEASUREMENT OF EDUCATIONAL COMPETENCE

An ad hoc committee was established by the Public Services Laboratory of Georgetown University July 1971 to: (1) review the concept of an adjusted educational achievement index, and (2) explore possibilities of applying a standardization ~~of population for~~ sex, income, race, and age (SIRA) differences to data collected on achievement test scores in comparing achievement scores among jurisdictions and across time. The committee met July 29 and September 25, 1971, with staff members of the National Center for Educational Statistics (NCES) and the Public Services Laboratory (PSL) of Georgetown University.

Members of the Ad Hoc Committee on Educational Achievement Measurement were:

Office of Education	Outside Consultants including PSL Staff
Dorothy Gilford	Alfred Carlson, Educational Testing Service
Boyd Ladd	William Coffman, Iowa Testing Programs
Ezra Glaser	H. Russell Cort, General Learning Corporation
Richard Berry	Burton R. Fisher, University of Wisconsin
William Dorfman	Virginia Herman, PSL, Georgetown University
	Selma J. Mushkin, PSL, Georgetown University
	Nelson Noggle, Science Research Associates

At the first meeting, committee members considered a preliminary statement of the need for a "SIR" adjustment, looked at technical and administrative problems, and discussed methods of adjusting for differences in population characteristics. The meeting ended with a tentative list of recommended steps.

The second meeting focused on the role of NCES in data collection and on needs for educational outcome measures. It discussed in greater detail methods of adjusting achievement and other competence scores. Finally, the committee agreed to recommend next steps that would: (a) produce a body of knowledge on what achievement tests are given to what children/by what jurisdictions, (b) provide information on how achievement test data are being used, and (c) seek to provide measures on outcome (through a SIRA-type of adjustment or otherwise) that would reduce possible misinterpretation.

The committee at its September 1971 meeting recommended:

1. That a survey be made on the extent of data on achievement scores. Study of State testing programs to update and elaborate the 1967 ETS compilation was proposed to determine what tests are being given to what children, where, and on what forms the data are reported. It may be feasible to tie a survey such as this into the "Longitudinal Study of a Representative Sample of the High School Class of 1972" now being conducted by NCES.[1] In the intervening period ETS has completed a new survey of State uses of educational achievement tests.(6)

2. That the data problems in SIRA adjustments be reviewed and analyzed. A proposed study of the information sources on SIRA by State and school districts would include the types of data, definitions, frequency of reporting, and application as a SIRA adjustment. Special attention should be given to the time lag between the need for decision-making or policy-making data and its actual availability.

3. That a small-scale test of SIRA be conducted. A study, in one or more pilot States, of SIRA adjusted achievement scores would ascertain: (a) the problems in gathering, reporting, and interpreting such scores, and (b) use of adjusted scores by States and districts.

4. That current uses of educational achievement scores as statistical materials be determined. Included in this study would be an in-depth examination of the current uses of achievement score data in each State. The study would build on the review suggested in Recommendation 1 and would determine the purposes of States and communities in using achievement test scores, for example interjurisdictional fund allocations, payment incentives and budgetary decisions within a government.

Dr. H. Russell Cort had reservations about making adjustments in achievement test data for evaluation. "For research purposes, it's certainly necessary and desirable . . . to be able to adjust groups for the purposes of comparison." But for purposes of direct practical decision-making, "the adjusting of test scores to take account of variations in population may, in fact, lead to either expectations or conclusions that are not desirable." Dr. Burton R. Fisher shared these reservations, although he believed that in selected instances SIRA adjustments may aid in decisions.

Dr. Cort concurred with Recommendations 1 and 2 but had reservations about 3 and 4. He said:

> The pilot project would be a very desirable thing although I find it difficult to reconcile even the notion of the pilot project with a value conviction on my part that once the Pandora's Box of adjustments is opened, the consequences of adjusted data are apt to be destructive or deleterious and ultimately beyond the control of the data provider. However, as we discussed the approach on September 25, 1971, it was agreed that, hopefully, a State that would be involved in trying out the use of adjusted data would have agreed to make use of it and hopefully would have agreed to explore in some detail the implications and actual effects of providing or publishing adjusted comparative data among school systems.

APPENDIX 2: CHARACTERISTICS OF SELECTED STATE-WIDE TESTING PROGRAMS*

A limited number of State education agencies were asked by the NCES in Spring 1973 for information on achievement test data representative of individual school districts.

Nine states were contacted: California, Florida, Iowa, Massachusetts, Michigan, Mississippi, New York, Texas, and Virginia. The findings reported for the nine states are summarized below:

California: Collects test data annually on the universe of students in grades 1, 2, 3, 6 and 12. The Cooperative Primary Reading Test is used at grades 1, 2 and 3. The California Test of Basic Skills is used at grade 6. The Iowa Test of Educational Development is used at grade 12. Distribution of scores by percentile by district could be made available.

Florida: Collects test data annually on the universe of 9th and 12th grade students (for purposes of high school program selection and for scholarship eligibility). In 1972, collected reading test data on a State-wide sample of 2nd and 4th grade students. Inferences can be made about districts, but the data are not technically district representative.

Iowa: Does not have an official statewide testing program. It is estimated that 425 of the State's 440 districts (over 90%) voluntarily test students on an annual basis using ITBS and ITED.** The State department does not have data. Release would require permission of each individual school district.

* Based on a June 14, 1973, memorandum to Selma Mushkin from Kathy Wallman.

** Iowa Tests of Basic Skills and Iowa Tests of Educational Development, respectively.

Massachusetts: Has no statewide testing program. Two years ago all 4th grade students were tested in a sample of 57 school systems, using the California Test of Basic Skills.

Michigan: Collects test data annually on the universe of students in grades 4 and 7. The Michigan Educational Assessment Battery, including Reading and Mathematics portions, is used. Deciles are available for all districts.

Mississippi: Has a voluntary statewide testing program at grades 5 and 8. Approximately 120 of 150 districts, or 85% of total enrollment, have participated. For those districts which have participated, it would be possible to obtain percentile distributions by district.

New York: Collects test data annually on the universe of students in grades 3 and 6. The New York State Reading and Mathematics tests are used. Stanines are available by district.

Texas: Has no statewide testing program.

Virginia: Collects test data annually on a universe of 4th, 6th, 9th and 11th grade students. SRA is used at the elementary level; STEP is used at the secondary level. Mean scores are available (published) for each district. Data are not stored in automated form; an investigator would have to use individual tests (hard copy) to analyze data further (e.g., to obtain percentile distributions by district).

Citations have been received by NCES suggesting that data on State testing programs may be available in the following additional States: Arizona, Delaware, Georgia, New Jersey, Nevada, and Wisconsin.

APPENDIX 3: BASIC TESTING PROGRAMS IN MAJOR SCHOOL SYSTEMS

City or County	Selected testing instruments by grade level						
	ITED[1]	ITBS[2]	STAT[3]	CAT[4]	STEP[5]	SRA[6]	MAT[7]
Akron[8]							
Albuquerque	9,11		5,7	5,7	2,3,4,6,7,8		
Ann Arundel, Md		3,4,5,6,8					
Atlanta							4,5,6,7
Baltimore City			5,6				3
Baltimore County		3,6,8			11		
Birmingham			4,5,6,7	8,11			
Boston							
Brevard, Fla			1,2,3,4,5,6,7 8,9,10,11,12				
Broward, Fla.				2,3,4,5,6			
Buffalo			4,6,8				
Caddo Parish, La	9			4,6,8,12			
Charlotte Mecklenburg			3,6,8,10				
Chicago	11						3,6,8
Cincinnati							
Clark, Nev			2	4,6,10,12			
Cleveland							
Columbus							
Dallas							
Dayton				3,5,6,8			
DeKalb, Ga							
Denver			1,2,3,4,5,6,7,10	8			
Detroit		4,6,8			10,12		
East Baton Rouge, La	10	8					4,5,6
El Paso	9	7		5			
Fairfax						4	
Flint						6,8	
Fort Worth		3,4,5,6	8				
Grand Rapids	10		1,2,3,4,5,6				
Greenville		3,4	4,6				
Honolulu					5,7,9,10,11,12		
Houston	9	3,4,5,6					
Indianapolis			6,8				4
Jacksonville			1,2,3,4,5,6				
Jefferson, Ala				8,11			
Jefferson, Col		3,5,8					
Jefferson, Ky			1,2,3,4,8,10				
Jersey City							
Kanawha			1,3,6				
Kansas City	10	4,5,6,8					
Long Beach	9	5		8			
Los Angeles	12			5,7,8			
L.A. County							
Louisville							1,2,3,4,5,6, 7,8,10,11
Memphis							1,2,3,4,5,6,7,8,10
Miami			2,3,4,5,6				7,8
Milwaukee		1,6,8					
Minneapolis	10	6,8					
Mobile							
Montgomery, Md.		5,7					

City or County	Selected testing instruments by grade level						
	ITED[1]	ITBS[2]	STAT[3]	CAT[4]	STEP[5]	SRA[6]	MAT[7]
Nashville Davidson							2,3,4,5,6,7,8,10,11
Newark							3,6
New Orleans	10	6			8		3,5
New York		4,6					
Norfolk					12	4	
Oakland			6				
Oklahoma City				6	8,10		3,4
Omaha		3,4,5,6,7					1,2
Orange Co., Fla							
Palm Beach			3,4,5,6,7,8				9
Philadelphia		3,4,5,6,7,8	2				
Phoenix	9,11		4,5,6,7,8	5,6,7,8			4,5,6,7,8
Pinellas, Fla.			4,6,7				
Pittsburgh							1,2,3,4,5,6,7,8
Portland							
Portland (Metro)							4,6
Prince Georges, Md.	9,11	5,6,7,8					3
Providence							
Richmond			3	6		4	
Rochester		4,6,7					
St. Louis	11	4,5,6,7,8					
St. Paul	9,11	4,5,6,7,8					1,3
San Antonio	11	3,5,6	3,4,5,6	7,8		7,11	
San Diego	10	6	8				
San Francisco			3,6,7,8,9				
San Jose							
Seattle							4,6,8,10
Syracuse	11	3,4,5,6,7,8					
Tampa							2,3,4,5,6,7
Toledo		3,4,6,8					
Tucson			5,8				3,4,6
Tulsa	9,11		3,4,5,7,8				
Washington							2
Wichita	10,11,12	3,4,5,6					
Worcester			3,6,8				
Youngstown							

[1] Iowa Tests of Educational Development

[2] Iowa Tests of Basic Skills

[3] Stanford Achievement Tests

[4] California Achievement Tests

[5] Sequential Tests of Education Progress

[6] Science Research Associates Achievement Series

[7] Metropolitan Achievement Tests

[8] In some of the school systems listed, other tests are used.

Source Akron Public Schools. "Basic Testing Programs Used in Major School Systems Throughout the United States." Akron, Ohio: Akron Public Schools, April 1968.