DOCUMENT RESUME

ED 393 687 SE 058 130

AUTHOR	Stecher, Brian M.; Koretz, Daniel
TITLE	Issues in Building an Indicator System for Mathematics and Science Education.
INSTITUTION	Rand Corp., Santa Monica, Calif.
SPONS AGENCY	National Science Foundation, Arlington, VA.
REPORT NO	DRU-467-NSF
PUB DATE	Jan 96
NOTE	107p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Educational Assessment; Elementary Secondary Education; *Mathematics Education; *Science Education; *Student Evaluation

ABSTRACT

In recent years, policymakers have shown renewed interest in the development of educational indicators hoping that specific quantitative indices will help them monitor the status of the educational system, understand its failures and successes, and build more effective remedies for perceived problems. This document is the final report of a study of the feasibility of developing a patchwork indicator system for science and mathematics education based on existing data sources covering achievement, secondary curriculum, and teacher workforce. This report summarizes the results of the patchwork exercise and offers recommendations for improving mathematics and science indicators. Contains 70 references. (MKR)

SE

ED 393 687

RAND

Issues in Building an Indicator System for Mathematics and Science Education

Brian M. Stecher and Daniel Koretz

DRU-467-NSF

January 1996

Prepared for the National Science Foundation

RAND

Issues in Building an Indicator System for Mathematics and Science Education

Brian M. Stecher and Daniel Koretz

DRU-467-NSF

January 1996

Prepared for the National Science Foundation

PREFACE

This is the final report of a study of the feasibility of developing a patchwork indicator system for science and mathematics education based on existing data sources. Previous reports described the quality of available indicators of achievement in mathematics and science, the quality of indicators of the high school curriculum in these subjects, and approaches to validating national indicators. This report summarizes the results of the patchwork exercise and offers recommendations for improving mathematics and science indicators. This document should be of interest to policymakers concerned about the status of the nation's mathematics and science education system and to those responsible for funding, designing or conducting national data collection efforts that describe mathematics and science education.

CONTENTS

TABLES

SUMMARY

In recent years, policymakers have shown renewed interest in the development of educational indicators. They hope that specific quantitative indices will help them monitor the status of the educational system, understand its failures and successes, and build more effective remedies for perceived problems. Researchers are more cautious about achieving these goals, raising both theoretical and practical concerns about the efficacy of indicators for these purposes. This research is designed to test the feasibility of a "patchwork" indicator system built from existing data sources, and to use this experience to inform the larger indicator debate. Our approach is both developmental and evaluative. We develop an indicator patchwork based on existing national data to describe three important features of mathematics and science education—teacher workforce, secondary curriculum, and student achievement. While constructing the indicators, we also evaluate the quality of the core data and the validity of alternative indicator definitions for answering policy questions, such as describing the educational experience of policy-relevant groups like racial and ethnic minorities or high-achieving students. The exercise provides helpful lessons about the problems of indicator system development and leads to recommendations regarding NSF's role as a sponsor and user of educational indicators.

THE CONDITION OF EXISTING INDICATOR DATA IN MATHEMATICS AND SCIENCE EDUCATION

A moderate amount of data is available to describe student achievement, curriculum, and the teacher workforce, but the collection does not provide a comprehensive picture of any of these areas. Furthermore, little or nothing can be said about the relationships among these three elements. This means that the patchwork of indicators drawn from existing data has serious limitations for many purposes. However, available data yield some useful information about achievement, secondary curriculum and the teacher workforce in mathematics and science.

Achievement

Although American students participate in a large volume of achievement testing every year, surprisingly little can be said with confidence about the achievement of the nation's students as a group. The reasons include inconsistent testing from one jurisdiction to another, apparently widespread corruption of test scores from teaching to the test, self-selection of students for certain tests (in particular, college-admissions testing), and insufficient information to reconcile discrepancies in findings across databases. However, several basic conclusions are warranted:

- The achievement of elementary and secondary students in many subjects, including mathematics and science, declined considerably during the 1960s and 1970s.
- The decline in achievement ended as cohorts born in the early 1960s moved through school, finally reaching the senior high level around 1980.
- Trends since 1980 are less clear. It appears that achievement has generally been increasing, but the size of the upturn relative to the preceding decline varies.
- The gap in test scores between African American and white students remains large but has narrowed markedly over the past few decades. Data about Hispanic students are sparser and less consistent but suggest that they have gained relative to non-Hispanic whites as well.
- Data about trends among high-achieving students, or students in the pipeline of future scientists and engineers, are inadequate and inconsistent.

Secondary curriculum

A detailed analysis of secondary curriculum at the classroom- and state-levels reveals a mixed picture with overall increases in graduation requirements and the availability of mathematics and science courses but major discrepancies in student course taking patterns among schools:

- There has been general improvement in students' exposure to mathematics and science over the past 15 years; between 1980 and 1985 graduation requirements increased markedly in both subjects and the number of mathematics and science courses completed by students has followed.
- However, differences in course completion rates among racial/ethnic groups are large, and they have been growing larger. In addition, group differences increase as the level of the course increases.
- Wide disparities remain between schools in some important aspects of curriculum; although basic and intermediate mathematics and science courses are available almost universally, advanced courses in both subjects are not available in 20%-25% of high schools.
- These disparities in course availability are associated with particular groups of students and particular school characteristics; advanced courses are less likely to be available to students in urban areas, in small schools and in schools with high minority populations
- Only a small percentage of students take comprehensive sequences of mathematics or science courses that are needed for mathematics and science careers; although these percentages are increasing, there are wide and growing disparities in favor of Asian and white students over African American and Hispanic students.

Teacher workforce

Existing data portray the demographic characteristics of the current teacher workforce rather well, provide moderate amounts of information about supply of and demand for new teachers, and illuminate some aspects of teachers' qualifications. RAND tabulations of the 1987-88 Schools and Staffing Survey (SASS) and other sources provide a somewhat disappointing picture of the teacher workforce:

- The overwhelming majority of mathematics and science teachers have the minimum qualifications necessary to teach in their subject field, but a substantial proportion are underqualified based on standards recommended by professional groups of mathematics and science educators.
- Qualification levels are higher for full-time math and science teachers and for teachers at the high school level than for part-time teachers and teachers in lower grades.
- There are continuing shortages of qualified mathematics and science teachers (regardless of which qualification standards are used), particularly teachers qualified to teach the more advanced and specialized science courses.
- The current teacher workforce in mathematics and science is divided relatively evenly between males and females, but it is overwhelming white, a situation which is unlikely to improve in the near term because an even smaller percentage of newly hired mathematics and science teachers come from minority population groups.
- The number of new teacher candidates being prepared by colleges and universities is declining, but the number entering teaching through alternative, non-traditional routes is increasing, and non-traditionally trained people account for a growing proportion of newly hired mathematics and science teachers. Unfortunately, this change makes it difficult to project the future supply of mathematics and science teachers.

LIMITATIONS OF THE CURRENT INDICATOR PATCHWORK

The patchwork of mathematics and science indicators that can be constructed from existing data is not sufficient to answer many of the questions currently posed by policymakers and the public. This mismatch between the information supplied by the current patchwork and that demanded of it can lead to invalid and misleading inferences. The insufficiency of the current patchwork derives from a number of factors, the most important of which is that the current array of data for the most part was not designed to provide a coherent indicator system. This

deficiency is apparent in the three domains we examined, and there is every reason to believe worse deficiencies exist in other, less-well-studied areas of mathematics and science education, such as student motivation, informal education, cooperative learning, parental support, etc.

One problem with the current patchwork is that coverage is incomplete, i.e., the patchwork does not describe many important aspects of mathematics and science education. There are two principal limitations to what is measured in the current patchwork: key constructs are measured only in very broad terms that reveal nothing about important details, and measures are operationalized in very traditional ways that are less than ideal for use in indicators. Both conditions limit the use of the patchwork for educational policymaking.

A second problem relates to measurement quality. Although researchers and practitioners generally have confidence in the accuracy of data gathered through federally supported data collection efforts, we find contradictions and limitations in current educational data that raise questions about their appropriateness for national indicators.

The third limitation has to do with verifiability; doubts about data quality linger because of the more general problem of too few sources on which to build and evaluate indicators. The quality of many data sources cannot be evaluated thoroughly due to a lack of secondary sources. Without periodic verification of this sort, national data collection efforts provide a basis of unknown adequacy for building indicators.

PRINCIPLES FOR DESIGNING AN IDEAL INDICATOR SYSTEM

The patchwork exercise reveals much about the characteristics of an ideal indicator system. For example, the thinness of the indicator patchwork can be attributed to national data collection efforts that were designed with a variety of different purposes in mind. If one were to start from scratch to build an indicator system for mathematics and science education, clarity of purpose would be a necessary condition. Similarly, a number of principles for indicator system design can be derived from our analysis of the current indicator patchwork:

- Establish clarity of purpose
- Select components that are comprehensive with respect to purposes
- Choose variables that are sufficient to span components and permit desired inferences
- Maintain a level of detail that is adequate to describe significant differences
- Employ data collection strategies and formats that assess the full range of constructs
- Engage in data collection that permits aggregation at appropriate levels vis-a-vis purposes
- Make the frequency of data collection inversely proportional to rate of change in the underlying construct
- Insure that measurement quality is adequate for desired inferences
- Utilize multiple independent sources of data
- Be sure that measures are sensitive to changes in the phenomena under study
- Develop analyses and presentations that reveal the underlying relationships.

CONSTRAINTS ON INDICATOR SYSTEMS

The previous discussion seems to suggest that an indicator system should encompasses a large, diverse set of constructs each measured by multiple variables at fine levels of detail, and that there should be redundant systems operating in parallel. However, there are theoretical reasons why this approach is not optimum and practical limitations that make this goal unattainable. Both kinds of constraints need to be understood before making decisions about national indicators for mathematics and science education.

On the theoretical level, there is fundamental tension between simplicity and comprehensiveness that is inherent in the definition of indicators. By design, indicators are simple statistics, but they are valued as a way to understand diverse, complex, dynamic systems. An

immediate challenge in developing indicator systems is to balance simplicity and comprehensiveness. A desire for completeness and explanatory power argues for increasing the number of variables that are included, the number of ways each is measured, and the level of detail of observations. However, indicator systems are valuable because they are limited, succinct and parsimonious. The purpose of indicators is to illuminate key elements of larger phenomena in a simple and concise manner, and this purpose precludes measuring comprehensively. One cannot achieve both goals; compromise is required.

On a more practical level, large, comprehensive indicator systems are expensive, and resources for their development are limited. Shavelson, et al. (1987) estimated the cost for a comprehensive independent national indicator system to be between $23 million and $34 million in 1987 dollars (not including state-level indicators such as the NAEP Trial State Assessment), and recent experience with NAEP suggests that this might be an underestimate. There is no indication that the National Science Foundation or the US. Department of Education is likely to fund an effort of this size. These fiscal limitations translate into fewer variables and more limited measurement strategies.

RECOMMENDATIONS FOR IMPROVING INDICATORS OF MATHEMATICS AND SCIENCE EDUCATION

In a previous study, RAND described five indicator system options NSF might adopt; here we recommend that NSF develop a hybrid "supplementary" system that combines features from two of those approaches. We suggest that NSF use its resources to supplement existing data collection efforts to obtain more complete data in areas of interest (the "piggyback" approach) and possibly commission some data collection on a regular basis to provide longitudinal measures (the "cyclical studies" approach) or other in-depth data that are not available through large-scale efforts. This approach involves analyzing existing efforts, identifying deficiencies, reforming data collection where possible to increase its utility, and creating new research when it is necessary in order to address issues not covered by available efforts and to test reliability and validity.

NSF must decide which roles it will play in promoting indicators and which roles it will leave to other agencies and research institutions. Our suggestion for NSF's role reflects assumptions about NSF's responsibilities for indicators and our view that the current indicator system is deficient and in need of research and development. We assume that NSF's role is more circumscribed than that of statistical agencies such as the National Center for Education Statistics (NCES). Of course, NSF is primarily interested only in mathematics and science education. Beyond that, NSF does not have responsibility for routine and cyclical data collection (such as the Common Core of Data maintained by NCES). The narrower purview of NSF provides an opportunity; lacking certain routine data collection responsibilities and free to concentrate on a few subject areas in depth, NSF has the prerogative to be more forward-looking in its approach to indicators.

Given these premises, we urge NSF to be the most forward-looking agency in the federal education indicator effort. It should leave to others most of the operational responsibilities for design, data collection, analysis, and reporting of routine data collection efforts. We believe that NSF should focus much of its efforts in two ways. First, it should support diverse supplementary data collection efforts, including add-ons to routine data collection and additional special studies. This is a traditional role for NSF reflected, for example, in its support for such efforts as TIMSS. Second, NSF should support planning, research and development, and evaluation pertaining to indicators. This latter focus would include, for example, experimental uses of new indicators, validation research, and periodic benchmarking studies.

We recommend that NSF develop an advisory infrastructure to guide its actions vis-à-vis indicator design and development. NSF should create a standing Indicator Advisory Group (IAG) with responsibility for monitoring its supplementary indicator efforts. The IAG should undertake tasks such as building consensus about purposes, evaluating existing data collection activities, establishing priorities for supplemental data collection, communicating with other agencies to increase the utility of their efforts, conceptualizing new studies that

would address issues not covered by available efforts or test the reliability and validity of existing data, and monitoring indicator-related efforts at the national level. The group should view its key responsibility as diagnosis and improvement, i.e., asking critical questions, identifying shortcomings, gaps and problems, and recommending actions to resolve them. Members should include researchers and representatives of relevant federal agencies and research organizations. We also suggest that the standing IAG be supplemented as appropriate with *ad hoc* committees with specific foci or expertise.

Finally, consistency of planning and funding is needed to overcome the irregularity and volatility of data that make a patchwork indicator system unstable. Therefore, it is important to maintain funding for the IAG and for key research activities for an extended period. Without continuity of funding, purpose and leadership, we will continue to have a haphazard patchwork rather than a useful indicator system. Even though NSF lacks line authority for much of the national education indicator effort, NSF can enhance the value of its indicator efforts by maintaining consistency in its planning, research and development, and data collection.

ACKNOWLEDGMENTS

The authors wish to thank Elizabeth Lewis and Lisa Hudson for their assistance with data analysis and Donna White and Judy Wood for their help with text preparation.

1. INTRODUCTION

In recent years, policymakers have shown renewed interest in the development of educational indicators.[1] Legislators and other policymakers hope that specific quantitative indices will help them monitor the status of the educational system, understand its failures and successes, and build more effective remedies for perceived problems. Researchers are more cautious about achieving these goals, raising both theoretical and practical concerns about the efficacy of indicators for these purposes. This project tests some of these concerns by attempting to develop a descriptive indicator system for mathematics and science education based on available data. The exercise provides helpful lessons about the problems of indicator system development and leads to recommendations regarding NSF's role as a sponsor and user of educational indicators.

CALLS FOR INDICATORS AS EDUCATIONAL POLICY TOOLS

The current interest in educational indicators is widespread, coming from legislators and government officials at both the federal and state levels as well as from educators and researchers. During the past few years there have been indicator-related initiatives emanating from the White House, the Congress, and the State Houses, including the National Educational Goals (National Educational Goals Panel, 1991, 1992), national curriculum standards and curriculum-related assessments (e.g., Goals 2000: Educate America Act), and school delivery standards. Furthermore, a growing number of governmental agencies and professional organizations are developing, collecting, and/or disseminating indicators of specific components of the educational system. A sampling of recent efforts includes indicators of the condition of proprietary schools (Goodwin, 1991), undergraduate education (Adelman, 1989), America's teachers (Choy, et al., 1993) and vocational education (Hoachlander, et al., 1992). It also includes indicators of the health

[1] The social indicator movement of the 1960s was similar in purpose and approach to the indicator movement of today (Shavelson 1987).

of education at the state-level (Blank and Gruebel, 1993; Blank and Dalkilic, 1990), the national-level (Special Study Panel on Education Indicators, 1991), and the international level (Lapointe, et al., 1992; Lazar, 1992). In the fields of mathematics and science education, the National Science Foundation is both a producer of indicators (National Science Board, 1993) and a source of funding for indicator research (Shavelson, et al., 1987; McDonnell, et al., 1990; Blank and Gruebel, 1993), including the present study.

The current interest in indicators is motivated by a variety of disparate goals. Some people would use indicators primarily to describe the status of education (Bracey, 1992; Bracey, 1993; Huelskamp, 1993), while others look to indicators for more complex purposes, including explaining educational phenomena, linking policies with outcomes, evaluating alternative programs, and predicting the effects of prospective actions. Those who subscribe to the latter goals think that indicators can be powerful policy tools for answering questions such as why students in one state achieve more on average than students in another, which science education programs are more effective, or how particular policy options will affect participation in mathematics and science courses. Many policymakers have high expectations that indicators can address diverse, complex questions such as these.

CAUTIONS ABOUT THE EFFICACY OF INDICATORS

Researchers are far more cautious than policymakers in their appraisal of the utility of educational indicators. They raise theoretical and practical questions about the construction of indicator systems and their use for the purposes of greatest interest to policymakers. From the research perspective, the development of high quality indicator systems is hindered by a number of factors, including an unclear conceptualization of indicators, inherent conflicts among policymakers' goals, and unrealistic expectations regarding the purposes that can be served by indicators. Furthermore, researchers point out that even for appropriate purposes the construction of an indicator system involves compromises that limit the usefulness and applicability of the data.

One obstacle to achieving policymakers' goals is an unclear conceptualization of indicators. Educational indicators are arbitrary statistically indices that are defined by policymakers and analysts, and an indicator system is a collection of these statistics, chosen to reflect elements that researchers or policymakers believe contribute to an understanding of the functioning of the educational system (Shavelson, et al., 1987; Oakes, 1986). Some policymakers tend to reify the statistics and act as if they are essential features of education. The inappropriate focus on mean SAT scores as a measure of educational success may be an example of this problem. It is possible to take metaphor of "measuring the health of the educational system" too seriously, assign to indicators greater credibility than is warranted.

A second problem arises because of the multiplicity of policy goals that might be served by indicators. An indicator system designed to serve one goal may not be effective (or even appropriate) for addressing another. Different purposes, such as describing the status of the current system, explaining relationships among educational components, predicting the effects of policy initiatives, or evaluating the quality of educational programs, require different kinds of information, and features that are important for one use may limit the value of the indicator system for other uses. For example, a system designed to describe the flow of students into mathematics and science careers will be quite different from a system whose purpose is to evaluate the relative merits of alternative science curricula. Similarly, a system that is optimized for describing trends in achievement will likely be inadequate for explaining those trends. Thus, the large number of potential goals is a hindrance to achieving any one of them.

A third caution arises because some of the goals policymakers hold f · indicator systems are unrealistic. In particular, causal questions dc ,nate the policy debate, and policymakers tend to use indicator data in attempts to explain successes and failures. However, indicator data are poorly suited to addressing causal questions. There have been many unsuccessful attempts over the years to try to use statistical indicators to understand social phenomena, and almost without exception, these attempts to use indicators to establish cause and effect

relationships have been unsuccessful (Shavelson, 1987). They fail for a number of reasons. We have an incomplete understanding of the functioning of the educational system, and our inability to offer comprehensive explanations for current phenomena hampers attempts to predict the effects of policy changes. More importantly, indicator systems by their very nature are not well suited to assessing cause and effect. The types of data they include are usually insufficient to establish causal relationships. To provide convincing causal explanations one would need data on all factors that potentially affect educational outcomes substantially and therefore offer alternative explanations of differences in performance. By its very nature an indicator system lacks the detail and depth of information to eliminate alternative explanations and support causal inferences. For example, indicator systems typically lack longitudinal records and include only limited data on non-educational factors known to influence outcomes greatly.

Consequently, an indicator system alone usually cannot provide answers to many important policy questions, such as: Do increased graduation requirements in science produce more scientists and engineers? or Does the use of cooperative group learning in mathematics and science increase the teamwork skills of high school graduates entering the technical workforce? Indicator systems, however, can identify problems that need investigating and can *suggest* explanations that need to be tested with other forms of data. There are circumstances in which indicator data can be useful for helping to evaluate explanations, but doing so requires both care and, more often than not, additional data. For example, in the early 1980s, some social critics attributed the end of the achievement decline (see Chapter 2) to social and political events at that time. However, a careful look at indicator data was sufficient to cast doubt on that explanation (or at least its sufficiency), because it showed that the end of the decline *in the earlier grades* preceded the policies in question. Indicator data, *in conjunction with other forms of data*, were also useful for evaluating a number of other common explanations of the achievement trends of that period (see Koretz, 1987).

Even for more appropriate goals, such as providing rich description of the educational system, there are limits to what an indicator system can accomplish. These limits derive in part from the need for simplicity that is inherent in the notion of indicators. Resource demands also create practical limitations on the scope of indicator systems. Indicators provide a simple, shorthand way of looking at complex systems. If they are too elaborate or complex, they lose their usefulness for these descriptive purposes. Designers of indicators systems must balance desires for breadth and comprehensiveness against the demands of simplicity. An indicator system that provides enough information to address a wide range of concerns and the full scope of the educational system will not offer the ease of access that is supposed to characterize indicators.

Similarly, resource limitations force developers to make compromises between opposing goals, such as breadth versus depth and cost versus quality. In the former case, one must strike a balance between an indicator system that covers a wide range of issues lightly or a system that covers a more limited set of issues in greater detail. In the latter case, there are similar tradeoffs between measuring a few features more accurately or a larger number of features less well. Decisions about such conflicting principles must be made (either explicitly or implicitly) when developing an indicator system, and each decision enhances the utility of the system for some purposes while constraining it for others.

LESSONS LEARNED DEVELOPING A PATCHWORK INDICATOR SYSTEM

This research was designed to test the feasibility of a "patchwork" system built from existing data sources (Shavelson, et al., 1987), and to use this experience to inform the larger indicator debate. Our approach was both developmental and evaluative. We developed an indicator patchwork based on existing national data to describe three important features of mathematics and science education--teacher workforce, secondary curriculum, and student achievement. While constructing the indicators, we also evaluated the quality of the core data and the validity of alternative indicator definitions for answering

policy questions, such as describing the educational experience of policy-relevant groups including racial and ethnic minorities or high achieving students (Stecher, 1992; Koretz, 1991; Koretz, 1992a). Chapter 2 presents a summary of these results, focusing on the status of these three features of the mathematics and science education system.

The study had a second purpose—to delineate key issues that must be addressed in any future indicator development. Our investigation of the patchwork approach to indicators helped us clarify the characteristics of an effective indicator system and identify many of the trade-offs that are inherent in developing such a system. Chapter 3 presents our findings with respect to indicator system development and the options NSF might pursue to maximize the impact of its contribution to indicators.

2. THE CONDITION OF EXISTING INDICATOR DATA IN MATHEMATICS AND SCIENCE EDUCATION

Available data support a "patchwork" of indicators describing the status of mathematics and science education, but this patchwork is incomplete and inadequate for many of the uses to which policymakers want to put indicators. Although it is possible to describe trends in student achievement, secondary curriculum, and the teacher workforce at a very general level with moderate confidence (Stecher, 1992; Koretz, 1993), building these indicators reveals a variety of deficiencies in current data sources. The delineation of these deficiencies can help to improve data sources and guide future indicator system development.

The chapter begins with a summary of the status of student achievement, secondary curriculum, and the teacher workforce based on current indicator data. The summary is followed by a discussion of some of the key limitations of the extant data as a basis for indicators. These problems affect the validity of inferences drawn from the data. The chapter closes with a few additional validity concerns that relate more to indicator systems than to specific data elements. The reader is reminded that this analysis was limited to three mathematics and science subdomains. Although many of the findings will generalize to mathematics and science education more broadly, the study should not be interpreted as a comprehensive investigation of all aspects of these domains. In the following chapter, we draw on our experience to recommend steps to improve indicators in mathematics and science.

KEY FINDINGS FROM CURRENT INDICATORS

A moderate amount of data is available to describe student achievement, curriculum, and the teacher workforce, but the collection does not provide a comprehensive picture of any of these areas. Furthermore, we can say little or nothing about the relationships among these three elements. This means that the patchwork of indicators drawn from existing data has serious limitations for many purposes.

This patchwork comprises a broad but unfocused body of data, including data collected with a descriptive purpose in mind and data

intended to answer specific research questions. The data include national surveys conducted specifically to provide indicator data (e.g., the National Assessment of Educational Progress); broad-based longitudinal surveys of selected age cohorts (National Longitudinal Survey of 1972, High School and Beyond, and the National Educational Longitudinal Survey of 1988); international comparative studies of curriculum and achievement (the Second IEA Study of Mathematics, the Second IEA Science Study, the International Assessment of Educational Progress); and individual studies of domestic educational policies (Education Commission of the States compilation of graduation requirements). In none of these three areas, however, do the available data represent a systematic strategy for describing all the important components of mathematics and science education. Instead, the "patchwork" of indicators that can be fashioned from existing data reflects enlightened historical happenstance, the result of a succession of individual research efforts and policy decisions, each with its own goals and focus.

In the following three sections we construct patchwork descriptions of the conditions of mathematics and science education in the areas of achievement, secondary curriculum and the teacher workforce. More detailed descriptions of achievement and secondary curriculum have been reported elsewhere (Koretz, 1991; Stecher, 1992). More detailed tables for the teacher workforce section are contained in Appendix A.

Achievement

Although American students participate in a large volume of achievement testing every year, surprisingly lit le can be said with confidence about the achievement of the nation's students as a group. The reasons include inconsistent testing from one jurisdiction to another, apparently widespread corruption of test scores from teaching to the test, self-selection of students for certain tests (in particular, college-admissions testing), inconsistent practices regarding the exclusion of special populations (such as students with disabilities or with limited proficiency in English), and insufficient information to reconcile discrepancies in findings across data sources.

Despite the limits of the available data, however, several basic conclusions are warranted:

- The achievement of elementary and secondary students in many subjects, including mathematics and science, declined considerably during the 1960s and 1970s. Declines in scores on most achievement tests, however, were generally smaller than the often cited drop in scores on the Scholastic Aptitude Test (SAT), because the latter was exacerbated by the lessening selectivity of the group of students taking the SAT.
- The decline in achievement ended as cohorts born in the early 1960s moved through school, finally reaching the senior high level around 1980.
- It appears that achievement has generally been increasing since about 1980, but the size of the upturn relative to the preceding decline varies substantially across databases, with many local and state databases showing greater gains than appear in the National Assessment of Educational Progress.
- The gap in test scores between African American and white students remains large but has narrowed markedly over the past few decades. Data about Hispanic students are sparser and less consistent, but some data suggest that certain Hispanic groups have also gained relative to non-Hispanic whites.
- Data about trends among high-achieving students, or students in the pipeline of future scientists and engineers, are inadequate and inconsistent.

In reaching these conclusions, we gave particular weight to the National Assessment of Educational Progress (NAEP), which is the only source of nationally representative, frequently collected data on the achievement of American elementary and secondary students. However, mindful of the risks inherent in over reliance on any single test, we also considered a wide variety of other measures of varying quality, including college-admissions tests, data from state-level testing programs, data from

national standardizations of commercial achievement tests, and data from infrequent large-scale nationally representative surveys.

Even the simple question of whether average achievement has declined or improved has been the source of considerable controversy in recent years. However, the patchwork of data suggest a number of conclusions:

Achievement, at least as measured by tests, clearly declined during the 1960s and 1970s. This decline was amply documented in mathematics, which is covered in one form or another in most assessment programs. Data about science achievement are much sparser; for example, most conventional commercial elementary and secondary achievement tests do not include science as a part of their core batteries, and many districts and states do not administer the optional science supplements. While numerous testing programs are instituting science assessments, data about science achievement remain less plentiful, and little of the science assessments are sufficiently long-standing to provide trend estimates. Nonetheless, as a whole, the science test data also suggest a decline.

The size of the decline in achievement — that is, the simple downward trend in mean scores — varied across data sets and remains a matter of considerable controversy, but when enough data are considered, it appears clear that the decline was sizable, particularly in the higher grades. Declines in college-admissions test scores (the SAT and ACT) were exaggerated by a lessening of the selectivity of the groups of students taking these tests, but even the drop in scores on tests unaffected by selectivity changes was often between 0.20 and 0.35 standard deviation and sometimes larger. For example, substantial declines in scores appeared in the National Assessment of Educational Progress, a comparison of equated results from the National Longitudinal Survey and the High School and Beyond survey, norming data from commercial test publishers, state-level assessment data, and a number of special studies. The exceptions were noteworthy — in particular, the American College Testing (ACT) science test, which showed a minor increase in scores — but few in number.

The size of the change in scores is shown in Table 2.1 for a number of databases that permit an estimate of the magnitude of the total decline; NAEP is excluded because its inception was too late to measure the early part of the decline. Mathematics and science data from the earliest administrations of NAEP are inconsistent in this respect. In mathematics (from 1972 through 1981), NAEP showed a moderate decline at age 17 but not at ages 9 and 13. (This pattern is consistent with cohort effect noted below, because the decline should have ended at age 9 by the mid-1970s.) In science (from 1069 through 1976), NAEP showed a more substantial decline at ages 9 and 13 but no consistent trend at age 17. At age 9, for example, the average percent of items answered correctly in the 9-year-old samples dropped from 61 to 52 percent during that period (Koretz, 1986, Table III-3).

Table 2.1

Magnitude of the Achievement Decline in Mathematics and Science, in Standard Deviations, Over Total Period of Measured Decline

Test and Subject	Grade	Decline
Mathematics		
Scholastic Aptitude Test	12	-.28
American College Testing	12	-.42
National Longitudinal Survey to High School and Beyond	12	-.14
Iowa Tests of Basic Skills U.S. standardization samples	12	-.26
Iowa Tests of Basic Skills U.S. standardization samples	10	-.32
Iowa Tests of Basic Skills U.S. standardization samples	8	-.28
Iowa Tests of Basic Skills U.S. standardization samples	6	-.28
California Achievement Test U.S. Standardization Sample	12	-.34
California Achievement Test U.S. Standardization Sample	9	-.30
Science		
American College Testing	12	+.06

SOURCE: Koretz, 1986.

The timing of the decline's onset is uncertain (because of limited and inconsistent data), but its end shows a reasonably clear *cohort effect*. That is, the decline ended, not in a particular calendar year,

but with particular birth cohorts, appearing later in the higher grades as the affected birth cohorts matured. Although the timing varies somewhat from one data source to another, the general pattern across a variety of data sources suggests that the decline's end typically occurred within a few years of the birth cohort of 1962, thus reaching the senior high grades—and public awareness—around 1980.

Although indicator data are generally poorly suited to explaining the phenomena they describe, this cohort pattern, in conjunction with the uncanny pervasiveness of the decline, may hold a key to explaining the trends. It is hard to explain a decline that appeared pervasively across jurisdictions, grades, and subject areas and that followed a cohort pattern in terms of specific changes in educational practice. Therefore, these patterns suggest that one or more societal factors contributed substantially to the decline and its end (Jencks, 1980; Koretz, 1987). Some specific societal factors can be identified; for example, it appears likely that demographic changes in the composition of the school-age population is accountable for a modest share of the decline in scores but impeded the more recent rise (e.g., Koretz, 1987).

It is clear that achievement in some subjects, including mathematics and science, has increased since the end of the achievement decline a decade and a half ago. (For a summary of NAEP trends in mathematics, science, reading, and writing, see Mullis, Dossey, Foertsch, Jones, and Gentile, 1991.) The evidence does not consistently show, however, that this improvement has fully offset the earlier decline. For example, NAEP documents trends in mathematics and science at three ages (9, 13, and 17). In mathematics at ages 9 and 13, performance was better in 1990 than in 1973, the first year tested, while in science at age 17, performance was considerably poorer in 1990 than in 1969, the first year tested (see Mullis, *et al.*, pp. 2-3, 1991). Moreover, because the NAEP assessments only began within a few years of 1970, they failed to capture part of the decline in scores and therefore overstate the size of the recent upturn relative to the size of the decline. Indeed, in mathematics, NAEP began too late (1973) to capture any of the decline among 9-year-olds and any but the end of the decline among 13-year-olds.

One reason that trends since 1980 are less clear than earlier trends is that growing pressure to raise test scores has apparently induced widespread inflation of test scores (Cannell, 1987; 1989; Koretz, Linn, Dunbar, and Shepard, 1991; Linn, Graue, and Sanders, 1990). This undermines the utility of secondary data sources, such as results of state and local testing programs, that ideally would be added to national data such as NAEP to build a patchwork of indicators. Indeed, Linn and Dunbar (1990) pointed out that recent trend data from state and local testing programs have shown sharper upturns in scores than has NAEP.

Over a span of several decades that included periods of both declining and rising average scores, African American students have gained relative to non-Hispanic white students. In some instances, these relative gains took the form of shallower declines among African American students, but during much of the period they comprised larger absolute gains by African Americans than by whites. This pattern appears with remarkable consistency across a variety of data sources. (See, for example, Burton and Jones, 1982; Koretz, 1986; Linn and Dunbar, 1990; Mullis, *et al.*, 1991, 1994; National Assessment of Educational Progress, 1981, 1985, 1988a, 1988b, 1990; for a direct comparison of data sources, see Koretz, 1986.) The relative gains of African Americans were in some cases quite rapid. For example, between 1973 and 1986, the gap on the NAEP mathematics assessment between African American and white 17-year-olds shrank by more than one-fourth. These relative gains are even more striking when one considers that the dropout rate among African-American students declined quite markedly between 1970 and 1985 or so (Koretz, 1990; National Center for Education Statistics, 1989a). A decrease in the dropout rate would be expected to depress test score gains by making the test-taking group less select. NAEP results from the past few years suggest that the relative gains of black students may be ending (for results in mathematics and science, see Mullis, *et al.*, 1994), but it would be wisest to accumulate several more years of data before reaching a conclusion about this.

Data pertaining to other minority groups are less clear-cut. Some data suggest that Hispanics also appear to have gained relative to non-

Hispanic whites (e.g., Koretz, 1986; Linn and Dunbar, 1990; Mullis, *et al.*, 1991), but this pattern is less consistent and striking than the gains of African Americans. Data about Hispanics are also affected by a number of limitations, including small samples, inconsistent classification rules, and a failure to distinguish among ethnically distinct Hispanic groups.[2] In addition, interpretation of trends among Hispanic students is clouded by rapid immigration. For example, improving achievement among native-born Hispanic students and immigrant students with long residence in the United States might be obscured by immigration of students with lesser proficiency in English and limited experience in American schools. Asian-American students often score higher than other groups on mathematics and science tests, but these data are subject to all of the limitations that affect data about Hispanic students and generally involve even smaller samples. For example, a recent summary of trends over two decades on the NAEP did not present data for Asian/Pacific Islander students because of insufficient sample sizes (Mullis, *et al.*, 1991, p. 25).

Curriculum

A detailed analysis of secondary curriculum at the classroom- and state-levels reveals a mixed picture, with overall increases in graduation requirements and the availability of mathematics and science courses but major discrepancies in course taking between schools (Stecher, 1992):

- There has been general improvement in students' exposure to mathematics and science over the past 15 years; between 1980 and 1985 graduation requirements increased markedly in both

[2]Unstable sampling of Hispanic subgroups has not been a problem for recent NAEP mathematics assessments. We found that although mean scores of the Hispanic subgroups differed considerably, the proportions of students reporting membership in each of the subgroups were quite stable from 1986 through 1992. Nonetheless, the possibility remains that trends within some databases and especially differences among them are confounded with different compositions of the groups classified as Hispanic.

subjects and the number of mathematics and science courses completed by students has followed.

- However, differences in course completion rates among racial/ethnic groups are large, and they have been growing larger. In addition, group differences increase as the level of the course increases.
- Wide disparities remain between schools in some important aspects of curriculum; although basic and intermediate mathematics and science courses are available almost universally, advanced courses in both subjects are not available in 20%-25% of high schools.
- These disparities in course availability are associated with particular groups of students and particular school characteristics; advanced courses are less likely to be available to students in urban areas, in small schools and in schools with high minority populations
- Only a small percentage of students take comprehensive sequences of mathematics or science courses that are needed for mathematics and science careers; although these percentages are increasing, there are wide and growing disparities in favor of Asian and white students over African American and Hispanic students.

For the purposes of this analysis, curriculum is defined broadly as those features of the educational environment that determine students' opportunities to learn. This includes features determined at the classroom level, such as the style and content of instruction, and features defined at the school, district or state-levels. Examples of curriculum elements that are more distant from instruction are course availability and graduate requirements. Available data are better for more distant curriculum elements: they provide a relatively complete description of graduation requirements, course availability, and course completion, but the data become quite limited for describing course content and instructional resources. We consider each of these aspects of curriculum in turn.

Graduation requirements increased during the first half of the 1980s, from an average of slightly less than one year each in mathematics and science to 2.1 years in mathematics and 1.8 years in science, reflecting a push toward high standards, but they have been relatively stable since that time. During this recent period of stability in graduation requirements, however, a small but increasing number of states created voluntary alternative graduation options leading to academically enriched diplomas. These enriched standards require approximately one additional year of coursework each in mathematics and science.

Course availability has changed little in the past two decades. Basic and intermediate college-preparatory courses in mathematics and science are available in almost all high schools, as they have been since at least the mid-1970s. However, advanced courses, such as calculus and physics, are unavailable in 20 to 25 percent of all high schools. Furthermore, there are substantial differences between schools in the availability of advanced mathematics and science courses. Advanced courses are less likely to be available to students if they attend urban schools with low parent-occupation profiles (i.e., a small percentage of parents employed in professional or managerial jobs and a large percentage of parents unemployed or on welfare), small schools, or high minority schools. Furthermore, these differences increase as the level of the course increases.

Course completion patterns have followed the same patterns as graduation requirements. As graduation requirements increased over the past few years, students completed more mathematics and science courses. The majority of students now complete a two-year core of basic and intermediate mathematics and science courses. Although the percentage of students that complete advanced courses is still relatively small, the percentage has been increasing rapidly. (See Tables 2.2 and 2.3)

Table 2.2

Mathematics Courses Completed

Course title (duration)*	Percent of high school graduates 1982	1987
Individual courses		
Any math	97	99
Any remedial math course/below grade level	33	25
Algebra I	65	76
Geometry	46	62
Algebra II (.5)	35	47
Trigonometry (.5)	12	19
Analysis/precalculus (.5)	6	13
Calculus (all)	5	6
AP calculus	2	3
Statistics/probability (.5)	0.3	0.4
Course combinations		
Algebra II + geometry	28	42
Algebra II + geometry + trigonometry	7	15
Algebra II + geometry + trigonometry + calculus	1	2

*All courses are at least one year in duration, unless otherwise noted.

SOURCE: Westat, Inc., 1988.

Table 2.3

Science Courses Completed

Course title (duration)*	Percent of high school graduates 1982	1987
Individual courses		
Any science	95	99
Biology	75	88
Chemistry	31	45
Physics	14	20
Engineering	0.1	0.1
Astronomy (.5)	1	1
Geology (.5)	14	15
AP/honors biology	7	3
AP/honors chemistry	3	3
AP/honors physics	1	2
Course combinations		
Biology + chemistry	28	43
Biology + chemistry + physics	11	17

*All courses are at least one year in duration, unless otherwise noted.

SOURCE: Westat, Inc., 1988.

Unfortunately, differences in course completion rates among population groups are large, and they have been growing larger. The proportion of Asian and white students who complete intermediate and advanced mathematics and science courses is much greater than the proportion of black and Hispanic students. Furthermore, these group differences increase as the level of the course increases. This effect is somewhat more pronounced in mathematics than in science. (See Tables 2.4 and 2.5)

Table 2.4

Combinations of Mathematics Courses Completed by Sex and Population Group

	Percent of 1987 high school graduates			
Groups	Never taken geometry	Geometry and algebra II	Geometry, algebra II, and trig-onometry	Geometry, algebra II, trig-onometry, and calculus
Sex				
Males	39	42	15	3
Females	38	43	14	2
Population group				
White	35	47	17	2
Black	56	29	8	1
Hispanic	60	24	7	2
Asian	19	62	31	15

SOURCE: RAND tabulations of 1987 High School Transcript Study.

Table 2.5

Combinations of Science Courses Completed by Sex and Population Group

	Percent of 1987 high school graduates			
Groups	Never taken biology	Biology I	Biology I and chemistry I	Biology I, chemistry I, and physics I
Sex				
Males	13	87	44	21
Females	10	90	42	13
Population group				
White	11	89	46	18
Black	14	86	29	9
Hispanic	15	85	28	8
Asian	8	92	66	42

SOURCE: RAND tabulations of 1987 High School Transcript Study.

In contrast, there are very few gender-related differences in course completion. With the exception of physics, in which males predominate, males and females complete most math and science courses in equal ratios.

Shifting the focus slightly from courses completed to enrollment (courses taken) permits us to examine the associations between course taking and selected school characteristics. Students are less likely to have taken intermediate mathematics and science courses if they attend urban schools with low parent-occupation profiles, schools with many students receiving subsidized lunch, small schools, and high minority schools. In general, the relationships between course-taking and school characteristics are stronger in mathematics than in science. Moreover, differences in course-taking are larger than the corresponding differences in course availability, and therefore cannot be explained completely by them.

Little is known about the actual content of mathematics and science courses or the manner in which topics are presented. Although a small number of textbooks dominate the market, this does not translate into a uniformity of content, because teachers do not "cover" the same proportion of the material in the books (Weiss, 1987). Non-representative studies of a few courses suggest that topic coverage differs markedly from class to class (McDonnell, et al., 1990), and that such difference are associated with differences in achievement (Westbury, 1992).

Finally, schools have acquired more laboratory facilities, computers, and calculators since the mid-1970s. However, although almost all schools have these kinds of instructional resources to some degree, they are not evenly distributed. Students in smaller schools and in urban schools with low parent occupation profiles are less likely to have access to computers, calculators, and specialized science laboratories of the type associated with advanced science courses. More importantly, there are wide variations in the use of such resources by teachers and students even when they are available.

Teacher Workforce

Existing data portray the demographic characteristics of the current teacher workforce rather well, provide moderate amounts of information about supply of and demand for new teachers, and illuminate some aspects of teachers' qualifications. RAND tabulations of the 1987-

88 Schools and Staffing Survey (SASS) and other sources provide a somewhat disappointing picture of the teacher workforce:[3]

- The overwhelming majority of mathematics and science teachers have the minimum qualifications necessary to teach in their subject field. However, a substantial proportion are underqualified based on standards recommended by professional groups of mathematics and science educators.
- Qualification levels are higher for full-time math and science teachers and for teachers at the high school level than for part-time teachers and teachers in lower grades.
- There are continuing shortages of qualified mathematics and science teachers (regardless of which qualification standards are used), particularly teachers qualified to teach the more advanced and specialized science courses.
- The current teacher workforce in mathematics and science is divided relatively evenly between males and females, but it is overwhelmingly non-Hispanic white, a situation which is unlikely to improve in the near term because an even smaller percentage of newly hired mathematics and science teachers come from minority population groups.
- The number of new teacher candidates being prepared by colleges and universities is declining, but the number entering teaching through alternative, non-traditional routes is increasing, and therefore non-traditionally trained people account for a growing proportion of newly hired mathematics and science teachers. Unfortunately, this change makes it difficult to project the future supply of mathematics and science teachers.

Teacher qualifications can be measured in a variety of ways, each of which produces a somewhat different impression of teachers' level of preparation. While the vast majority of mathematics and science teachers meet state certification standards, far fewer have college

[3]Selected tables are reported in the text. Additional data relating to the teacher workforce are reported in Appendix A.

majors in their subject, and fewer still meet standards adopted by professional organizations. The data indicate that approximately 85 to 90 percent of secondary-school teachers whose primary or secondary assignment is mathematics and science have the minimum qualifications required to teach their subject, i.e., they are fully certified to teach mathematics or science. This estimate does not include elementary teachers or secondary-school teachers who teach mathematics or science only occasionally. Fortunately, full-time teachers are more likely to be credentialed in mathematics or science than part-time teachers. Furthermore, this trend has less of an overall impact in mathematics than in science because only a small percentage of secondary-school mathematics teachers are "occasional" teachers. However, over one-third of secondary-school teachers in some science fields are "out of field" teachers who are less well qualified. (See Tables 2.6, A.1 and A.2.)

Table 2.6

Certification Status, Public Secondary School Teachers Whose Primary or Secondary Assignment is Mathematics or Science

Subject and grade level	Percent of teachers fully certified in subject*	Percent of teachers initially certified in subject+	Percent of teachers excluded from tally**
Mathematics			
Grades 7-8	89	93	6
Grades 9-12	92	96	7
Biological science			
Grades 7-8	91	94	42
Grades 9-12	94	97	10
Physical science			
Grades 7-8	82	87	38
Grades 9-12	87	92	29
General science			
Grades 7-8	87	89	8
Grades 9-12	86	92	19

*Fully certified includes teachers with "regular or standard" state certificates and teachers with probationary certificates, defined as "the initial certificate issued after satisfying all requirements except the completing of a probationary period."

+Initially certified includes temporary or emergency certification in addition to the other types.

**This column indicates the percentage of mathematics or science teachers of each type who taught the subject only occasionally and were excluded from the other tabulations

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Certification is not be the best measure of teacher qualifications; it is not standardized across states and does not reflect a very high level of preparation. For these reasons, college degrees, college major fields of study, and college coursework may provide better indicators of teachers' qualifications. Not surprisingly, these indicators suggests more frequent underqualification than does certification status alone. Substantial percentages of secondary-school mathematics and science teachers lack the type of college preparation deemed appropriate by professional groups of mathematics and science educators. For example, although almost all secondary-school mathematics and science teachers hold Bachelor's degrees or higher, approximately three-quarters of

junior high school mathematics teachers and one-half of high school mathematics teachers did not major in mathematics. Secondary-school science teachers were somewhat better prepared, but fully one-half of junior high school science teachers and one-quarter of high school science teachers did not major in science. Including teachers whose major was in mathematics education or science education in these tabulations increases the percentages considerably, but at least one-quarter the secondary-school mathematics teachers and one-sixth of the secondary-school science teachers did not major in their subject or subject education. (See Table 2.9 and Appendix A.)

Table 2.7

Percentage of Secondary-School Teachers with a Major or Minor in Mathematics or Science, in Mathematics or Science Education, or in Either Field

	Subject	Subject education	Either
Public school teachers			
Mathematics			
Grades 7-8	30	29	55
Grades 9-12	50	36	76
Science			
Grades 7-8	50	22	64
Grades 9-12	73	27	86
Private school teachers			
Mathematics			
Grades 7-8	25	15	37
Grades 9-12	53	24	68
Science			
Grades 7-8	45	18	52
Grades 9-12	80	10	85

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Underqualification is yet more prevalent if we define qualifications in terms of specific college-level courses in mathematics and science. Professional organizations in mathematics and science (National Council of Teachers of Mathematics, National Science Teachers' Association) have adopted standards for junior high school and senior high school teacher preparation that are defined in terms of coursework. About one-third of the current junior high school mathematics and

science teachers and about 40% of the current high school science and mathematics teachers fail to meet these course-based qualification standards (Weiss, 1987). The stricter one sets course-taking criteria the lower the percentage of teachers who are qualified.

Inservice training is another element of teacher preparation that affects a significant proportion of mathematics and science teachers. Over one-half of the secondary-school mathematics and science teachers surveyed in 1985 engaged in some form of staff development during the previous year. However, much of this training was short-term (less than 30 hours of study) (Weiss, 1987). Furthermore, only a small fraction of the extended continuing education taken by secondary-school mathematics and science teachers (30 hours of study or more) was in the fields of mathematics or science. Thus, the overall impact of inservice training on secondary-school teachers' subject matter preparation in mathematics and science was presumably small.

The pattern of underqualification described above is not uniform across grade levels or subject fields; there are consistent differences in teacher qualifications between junior and senior high schools and between specializations within science and mathematics. Senior high school mathematics and science teachers are more qualified on almost all measures than junior high school teachers. Similarly, secondary-school biological science teachers and mathematics teachers are generally more qualified than general science teachers and physical science teachers. Secondary-school general science teachers are the least well prepared in terms of college majors and college coursework. In addition, there are substantial differences in teacher preparation within the physical sciences at the secondary level. Secondary-school chemistry teachers are better qualified in terms of degree major and coursework than physics teachers, who, in turn are more qualified than earth science teachers. (See Tables A.3 - A.10.)

The secondary-school teacher workforce is well representative of the population in terms of gender, but considerably less so in terms of race and ethnicity. Small gender difference and large racial/ethnic differences will be found among secondary-school mathematics and science teachers. Females comprise about one-half of public school mathematics

and science teachers in grades 7-8, and slightly less than one-half in grades 9-12. The percentage of female secondary-school mathematics and science teachers is highest in mathematics, and lowest in science, particularly high school science. The percentage of females is 10 to 20 points higher in private secondary-school than in public secondary-schools. (See Table 2.8.)

Table 2.8

Demographic Composition of Secondary-School Mathematics and Science Teachers, by Grade Level

Subject and grade level	Percent female	Percent minority	Percent non-Asian minority
Public school teachers			
Mathematics			
Grades 7-8	55	13	12
Grades 9-12	49	8	7
Biological sciences			
Grades 7-8	48	9	9
Grades 9-12	41	8	7
Physical sciences			
Grades 7-8	49	9	8
Grades 9-12	34	9	9
General science			
Grades 7-8	45	7	7
Grades 9-12	43	12	10
Private school teachers			
Mathematics			
Grades 7-8	72	7	7
Grades 9-12	59	6	4
Biological sciences			
Grades 7-8	74	5	4
Grades 9-12	59	8	2
Physical sciences			
Grades 7-8	73	4	5
Grades 9-12	37	8	4

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Teachers from minority groups make up roughly 10 percent of the mathematics and science teachers in public secondary schools. Slightly larger percentages of minority group members will be found among junior high school mathematics teachers and high school general science teachers. Minority teachers comprise a slightly larger share of mathematics and science teachers than of the total teaching workforce. Recently hired teachers are more likely to be female but less likely to come from minority population groups than the overall current secondary-school teacher workforce, indicating the group differences may be increasing rather than decreasing. (See Tables A.11 - A.12)

There appears to be a continuing shortage of qualified secondary-school mathematics and science teachers. Over one-half of high school principals (public and private) reported difficulties hiring fully-qualified teachers in physics, chemistry, computer science and mathematics (Weiss, 1987). Shortages have been reported by college placement offices in these same subjects for over a decade (Moody and Christoff, 1992). There are "considerable shortages" of secondary-school teachers in Physics, Mathematics, Computer Science, and Chemistry, and "some shortages" in Computer Science, Earth Science, General Science and Biology (Nicholas, 1992).

Misassignment is one consequence of the shortage of qualified secondary-school mathematics and science teachers.[4] Here the term misassignment refers to the practice of assigning teachers qualified in one subject field to teach in another in which they are less well qualified.[5] Almost 20 percent of high school mathematics teachers report that they are not best-qualified or even second-best qualified to teach mathematics (i.e., mathematics is at best their third strongest subject field). Comparable figures from science range from 15 percent (high school biology) to about 60 percent (high school general science and junior high school earth science). (See Table 2.9.) Furthermore, about one-half of the secondary-school teachers who taught primarily in the physical sciences had switched from another primary assignment, and many of these came from non-science disciplines. For example, more than one-half of the secondary school teachers who switched into earth science had previously taught in non-science fields. About one-quarter of secondary mathematics teachers who switched into mathematics had previously been assigned to science fields and three-quarters had assignments outside of mathematics and science. (See Tables A.)13 - A.15.)

---

[4]Teachers may be misassigned to teach mathematics and science for many reasons, including work rules and small school size, but lack of qualified mathematics and science teachers is a major factor.

[5] The SASS data do not indicate whether the teachers were certified in the other subject.

Table 2.9

Percentage of Public Secondary School Mathematics and Science Teachers Who State that They are Best or Second Best Qualified to Teach in Their Subject Area

Subject and grade level	Best qualified field	First or second best qualified field
Mathematics		
Grades 7-8	67	80
Grades 9-12	76	84
Biological science		
Grades 7-8	46	59
Grades 9-12	74	84
General science		
Grades 7-8	33	58
Grades 9-12	19	41
Chemistry		
Grades 9-12	56	75
Physics		
Grades 9-12	26	57
Earth science		
Grades 7-8	23	39
Grades 9-12	33	53

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

It is difficult to estimate the future supply of mathematics and science teachers. Colleges are the main source of new teachers, and teacher preparation levels in colleges and universities have fallen to an all time low (Moody and Christoff, 1992). This bodes poorly for the supply of mathematics and science teachers in the future. However, the primacy of colleges of education as the supplier of new teachers has been declining in the past few years, and alternative routes to certification have expanded to ease entry into teaching for those with other occupational experience. Approximately 60 percent of newly-hired, inexperienced secondary-school mathematics and science teachers came directly from college, but 40 percent held other jobs in the year prior to taking their teaching position. Little is known about the qualifications of those hired to teach mathematics and science through non-traditional routes. By and large they are not people changing from

other mathematics and science careers; the bulk of those mathematics and science teachers who worked outside of teaching before taking their present job came from sales, administration, and administrative support. (See Tables A.16 and A.17.)

WHAT CANNOT BE SAID: LIMITATIONS OF THE CURRENT INDICATOR PATCHWORK

The patchwork of mathematics and science indicators that can be constructed from existing data is not sufficient to answer many of the questions currently posed by policymakers and the public. This mismatch between the information supplied by the current patchwork and that demanded of it can lead to invalid and misleading inferences.

Part of this mismatch has resulted from changing demands for information. The information needed for the education policy debate has increased rapidly and changed in character in recent years, and even a well-designed indicator system would be hard pressed to keep pace. For example, until the end of the 1980s, the policy debate at the national level rarely focused on differences among states in educational achievement, and no effort was made to provide representative data on achievement at the state level. Within the space of a few years, comparisons among states became a major focus of debate, and large amounts of money have been directed to the NAEP Trial State Assessment (the supplementary NAEP samples that provide state-representative estimates) in the past few years. Similarly, the current wave of educational reform, in contrast to the education reform movement of the 1980s, is focusing attention much more on the quality of curriculum and instruction rather than simple course requirements, and existing data on curriculum are therefore less adequate for current purposes than they were a few years ago.[6]

The insufficiency of the current patchwork, however, reflects far more than these changes in demand. The current array of data was for

[6]In addition, both policymakers and the public often turn to indicator data for causal inferences about systems—e.g., for information about which state education systems are more effective. In this section, we consider only the ways in which the current patchwork is insufficient for legitimate uses of indicators. Making the patchwork adequate for causal inferences would require very different and often competing changes in the current data system.

the most part not designed to provide a coherent indicator system. One problem with the current patchwork is that coverage is incomplete, i.e., the patchwork does not describe many important aspects of mathematics and science education. This deficiency was apparent in the three domains we examined—achievement, secondary curriculum, and the teacher workforce—and there is every reason to believe worse deficiencies exist in other, less-well-studied areas of mathematics and science education, such as student motivation, informal education, cooperative learning, parental support, etc.

Second, in areas in which the current patchwork offers coverage, it is of uneven quality. Some variables are measured well, but others are not. For example, for two decades, NAEP has been used to provide information about student achievement in "disadvantaged urban" communities. The classification of communities, however, is questionable on several grounds and has not been validated. For example, it relies on principals' estimates of the occupations of the parents of the schools' students, and there is as yet no data indicating that principals can estimate that accurately. In other cases, we found contradictions between available data, which may reflect the low quality of one or more data sources.

Third, in some instances, the patchwork must rely on too few high quality sources—sometimes a single source—for important information. This sparseness of sources presents the user of patchwork data with two unpalatable choices. One is to use the data from the single best source, or perhaps the few best. An example is the frequent use of NAEP data, without reference to any other, to portray patterns of achievement. Unfortunately, this can lead the user to unwarranted and misleading conclusions, because data sources often yield inconsistent findings.[7] Thus, users who rely on a single source may mistakenly

[7]A variety of factors can cause this, including differences in sampling, the phrasing of questions, the selection of test items, the scaling of results, and other aspects of instrumentation, but in practice it is often unclear why two sources differ. For example, differences in results among achievement tests are not rare (e.g., Berends and Koretz, forthcoming; Koretz, 1986), and even minor differences in test administration may substantially alter basic

accept idiosyncratic findings as robust and valid. The alternative to relying on a single source is to use secondary and tertiary data sources that have substantial weaknesses. This poses risks as well because secondary data sources are often of lower quality and may have flaws that make them misleading.

Coverage: What Is Measured?

There are two principal limitations to what is measured in the current patchwork: key constructs are measured only in very broad terms that reveal nothing about certain important details, and measures are operationalized in very traditional ways that are less than ideal for use in indicators. Both conditions limit the use of the patchwork for educational policymaking.

The current patchwork consists primarily of information about mathematics and science education defined in broad conceptual terms; it provides less information about the components of these constructs that are important for policy analysis. In all three areas we investigated the patchwork lacks policy-relevant details, such as the status of important subgroups of students or curriculum topics. The lack of details precludes the use of the patchwork for many comparisons of interest.

For example, data exist to describe access to mathematics and science curriculum at the course level and to monitor changes in availability of courses over time. However, there is little or no information about the content of courses, making it difficult to interpret the broader information on availability. This can lead to inferences whose validity is uncertain. On the surface, it appears that students have adequate access to Algebra courses, however, it is impossible to tell whether courses bearing this title cover the same content and provide equivalent preparation for further study. Recent research on the opportunity to learn specific topics in Algebra suggests that Algebra content varies considerably from class to class (McDonnell, et al., 1990). Furthermore, there is some evidence that schools were

---

findings (Beaton and Zwick, 1990; National Center for Education Statistics, 1989b).

"watering down" the content of Algebra I to make it accessible to more students (Clune, 1989). If this were true, the change in course completion might not be accompanied by an increase in exposure to basic algebra skills. Furthermore, many policymakers assume that an increase in the proportion of students taking Algebra I and other core academic courses will predict increased student success in later high school and college courses (as suggested by Pelavin and Kane, 1990), but this may not be so, and the current patchwork does not provide data to test this contention.[8] Similarly, the lack of information about course content limits the value of the patchwork for other uses, such as tracking the impact of curriculum reform efforts in mathematics.

National data on student achievement have similar limitations; they describe best the performance of a few types of students in broad content domains. They do not provide as much detail about narrower slices of content or selected subgroups of students. For example, the NAEP provides data on performance in mathematics for students in grades 4, 8 and 12. However, the range of questions asked is too limited to draw conclusions about specific subfields of mathematics and, although the sample of students is adequate to provide information about major population subgroups, it is too limited to support inferences about many other groups of students, such as high-achieving minority students. In addition, the sample used for trend estimates is smaller than that used for the main NAEP assessments, so estimates of relative trends among racial/ethnic groups-among the most important estimates produced by NAEP-are highly error-prone (Barron and Koretz, forthcoming).

Information about student performance in mathematics or science subfields would be useful for judging the impact of curriculum reform efforts. Similarly, information about the performance of population subgroups is relevant to questions about educational equity, and data on the performance of high and low achieving students is important for monitoring educational productivity, e.g., the pipeline of students likely to enter mathematics and science careers. The current patchwork does not adequately support such subgroup and subtopic indicators.

[8] The association between core coursework and later success found by Pelavin and Kane (1990) may reflect selectivity bias, for example.

Likewise, teacher qualifications can be monitored only in the broadest terms. For example, although there is a growing amount of data on the formal preparation of teachers, there is still no information available about teachers' instructional behaviors. We lack direct measures of skill, teaching methods, enthusiasm, sensitivity or motivation, all factors that affect instructional quality and define, in vivid terms, what distinguishes a good from a bad teacher.

These are examples of under-representation of important elements of a construct in an indicator. When this occurs, the indictor becomes a poor proxy for the construct of interest. The potential for such construct under-representation is greater when indicators are developed from existing data sources that were constructed for other purposes, as is the case in a "patchwork" model of indicators.

The second difficulty with coverage is that the current data collection efforts are limited to "traditional" operational definitions of key constructs—achievement, curriculum, teacher preparation, etc.—rather than more contemporary conceptions of these constructs. For example, existing national test-based measures of achievement are formulated primarily in terms of knowledge and still rely primarily upon multiple-choice items in which students select the best answer. Mathematics and science educators have been calling for achievement to be judged on the basis of demonstrated performance on realistic tasks, and measurement specialists have been pursuing this reform agenda. However, little of this change in apparent in the current patchwork.[9] On the other hand, performance-based assessments impose many costs (e.g., less coverage and less reliability per unit of testing time), and it remains to be shown whether more performance-based tests will be practical for indicator purposes and will provide information valid for those uses.

[9]The National Assessment of Educational Progress has begun introducing constructed response elements into many of the assessments. For example, open-response pencil-and-paper items have been included in the mathematics assessments for several years, and the hands-on assessments in science will be included in the main assessment for the first time 1995-96.

Similarly, course-based indicators of curriculum do not reflect the dramatic changes that are occurring in the way mathematics and science are structured and presented (NCTM, 1989; National Research Council, 1994). Current models for mathematics and science curriculum organize content in more interdisciplinary ways and place greater emphasis on cross-cutting, conceptual themes rather than discrete factual units. Instructional reforms are based on the notion that students should be active constructor of understanding rather than passive receptors of information. Current measures of curriculum reflect none of these changes.

Teacher workforce indicators also are rooted in old conceptions of teacher preparation and the teacher's role. Existing measures of teacher preparation are framed in terms of formal preservice coursework and certification. However, a growing proportion of new teachers enter the profession through alternative certification routes, and the existing measures are of limited relevance to their capabilities and training. Furthermore, efforts to change the governance and organization of schools—to decentralize authority, to base accountability on outcomes, and to give local constituents greater choices—will be accompanied by changes in the roles and responsibility of teachers. This may require reconceptualization of the qualities of effective teachers in ways that challenge existing data sources.

The problem of traditional operational definitions is not unique to indicators. Most large-scale educational data collection efforts tend to be traditional in conception. There are many reasons for this, including a desire for continuity with previous data collection (to permit the monitoring of trends), the negotiated nature of large data collection enterprises which puts a premium on commonly shared concepts, a reasonable reluctance to adopt a perspective that reflects short-lived fads rather than lasting changes, and the lengthy development process that can separate by years survey design from data availability. Secondary users of these data, including indicator developers, in some ways benefit from this conservatism but also must pay a price for it.

Measurement Quality: How Well Is It Measured?

Although researchers and practitioners generally have confidence in the accuracy of data gathered through federally supported data collection efforts, we found contradictions and limitations in current educational data that raise questions about their appropriateness for national indicators.[10] The most serious doubts arise when comparable data sources contradict one another. For example, NAEP and NSSME disagreed by 25 percentage points in the percentage of schools offering Trigonometry in 1985-86 (Stecher, 1992).[11] Similarly, data from NAEP and from commercial testing programs do not always describe similar trends in achievement. Indeed, various NAEP assessments occasionally appear inconsistent with each other; for example, the main (cross-sectional and short-term trend) assessment and the long-term trend assessment suggest very different conclusions about the extent to which eighth and twelfth-grade students differ in terms of writing proficiency. Such direct contradictions, although relatively uncommon, raise serious questions about the trustworthiness of the underlying data.

Far more common are questions about quality that derive from specific operational choices that are made in designing the data collection efforts and from procedural problems with existing data sources. For example, operational choices have created noncomparable definitions of locality. NAEP's characterization of communities (Size and Type of Community) does not match the definition of urbanicity used by most other surveys. Therefore, it is difficult to draw comparisons between NAEP results and other data about regional differences in schools. Similarly, NAEP relies in large measure on student reports for social background data, even though such measures are known to be of low quality for certain variables, particularly at young ages (Berends and

[10]We included annual or biennial surveys (such as the School and Staffing Surveys), longitudinal studies of school-aged youth (such as High School and Beyond, and the National Education Longitudinal Study), and special educational studies (such as the National Assessment of Educational Progress, and the International Mathematics and Science studies).

[11]However, they agree in general on the availability of most other courses.

Koretz, forthcoming; Fetters, Stowe, and Owings, 1984; Kaufman and Rasinski, 1991). Some of the instances in which we found contradictory information may reflect the low quality of one or more data sources.

Procedural problems, such as missing data or incomparable samples, are not uncommon in large-scale data collection efforts, and they can affect the quality of the data that are produced. For example, between 10 and 20 percent of the schools in the 1985-86 NAEP failed to report information on course availability. Similarly, the non-response rates from private schools on the three components of the 1987-88 Schools and Staffing Survey ranged from 21 percent to 34 percent (Choy, et al., 1993). This level of missing information leads to questions about the representativeness of the data. Furthermore, the findings highlight the care with which existing data sources must be evaluated. Such limitations are not apparent from a cursory glances at project reports. They may only be revealed by a careful, and skeptical, appraisal of the data. Although such operational choices and procedural problems are not *prima facie* evidence of erroneous data, they are examples of situations that raise doubts about data quality.

These questions of measurement quality arise in part because practical considerations lead to the use of simple and inexpensive strategies in large-scale data collection. Simple indices are appealing because of their ease of collection and apparent clarity, but if they fail to portray real differences accurately they lead to impressions that are incorrect. For example, a simple index of exposure to curriculum that has been used frequently in large-scale surveys is the number of years of mathematics coursework completed by a student. This simple measure can be gathered from a student in a single question, and on the surface it reveals much about the student's mathematics education. However, the measure obscures important differences between mathematics courses; three years of basic mathematics does not impart the same knowledge and skills as three years of college preparatory mathematics. Ignoring these differences creates an inaccurate impression about students' experience in mathematics.

Cost considerations also can affect the utility of measures for indicators. Although cost is an unavoidable consideration in research

design, cost-cutting can lead to the use of data collection methods that provide less accurate and less interpretable information. Student course-taking provides a good example of this concern. The most reliable way to determine which courses students have completed is to review student transcripts. These documents contain the official designations of course credits and grades. However, transcript reviews are expensive, so alternative methods of estimating course-taking patterns are frequently used. The most common method is to survey students about their own course-taking. Such self-reports are not perfect substitutes for transcripts, and there is some evidence of selective bias in reporting in certain subjects (Valiga, 1986). Self-reports are also used in many places were observations or surveys of parents would provide more accurate, complete, and trustworthy data. The tendency to use simple and inexpensive measures also leads to some of the coverage problems describe previously.

Verifiability: How Many Independent Sources of Information Exist?

Some doubts about data quality linger because of the more general problem of too few sources on which to build and evaluate indicators. For example, one cannot know whether the results of a survey are idiosyncratic if there are no other data sources against which to compare them. Similarly, some aspects of data quality, such as the seriousness of low response rates, may be difficult to gauge without additional sources of data. In the late 1980s two nationally representative surveys examined teachers' inservice training experiences, the 1985-86 National Study of Science and Mathematics Education (NSSME) and the 1987-88 School and Staffing Survey (SASS). Comparisons between the surveys provided validation for both. Although SASS continues to gather such data biennially, since 1987 there has been no second source of data that can be used to judge the accuracy and meaningfulness of the SASS results. The quality of many data sources cannot be evaluated thoroughly due to this lack of secondary sources. Without periodic verification, national data collection efforts provide a basis of unknown reliability for building indicators.

Despite the importance of multiple data sources for validation (and to help with interpretation of inconsistencies among sources), it is rare to find a construct for which there are multiple high-quality sources of nationally-representative data. For some constructs we have corroborating evidence from second sources with which to test our interpretations of indicators; for some we have a credible first source but lack a second of sufficient quality; for others we have only a first source.

Other Factors Affecting the Validity of Indicator Systems.

In an earlier project report, Koretz (1992) identified factors that affect the validity of inferences drawn from indicator systems. Some of these factors, particular those that operate at the level of an indicator *system* rather than the level of individual indicators, were not made apparent by the examination of the current indicator patchwork, but they nonetheless warrant consideration when thinking about improvements to the present system. At the level of individual indicators these threats to validity include the tendency to over-generalize, the corruptibility of measures, and the drift of constructs over time and context. The validity of inference made from indicators systems are threatened when the system is insufficiently broad to measure full policy impact and when it lacks data to test alternative explanations.

One of the most serious threats to the validity of inferences based on indicators is over-generalization, and it is a problem that is inherent in the nature of indicators. Indicators are simplified indices used to draw inferences about broad constructs, and this generality alone makes them susceptible to misinterpretation. As the breadth of the concept spanned by the indicator increases (e.g., from "two-place addition of integers" to "fundamental operations" to "arithmetic" to "computational skill" to "mathematics achievement"), the sensitivity of the index to lower-level differences diminishes. Such highly aggregated measures contain little or no information about the individual components from which they were constructed, so they create incomplete impressions about the status of the underlying variables which can lead

to invalid interpretations. Although this appears to be similar to the problem of inadequate coverage described earlier, it is quite different. It is possible to have adequate coverage of a domain, but to encounter problems with over-generalization from an indicator in that domain because a large number of measures are combined during the construction of the indicator and the information from the original measures is lost to users of the indicator.

A second validity concern is that indicators are often built from measures that are themselves corruptible, i.e., the value of the measure can be influenced by the manner in which it is used independent of changes in the underlying construct. A timely example of this can be drawn from recent research on the use of standardized test scores. When tests are used in low-stakes context (as in the case of NAEP, whose scores are not associated with individual students, teachers or schools) the results are likely to reflect the knowledge and abilities of students.[12] However, when tests are used in high-stakes contexts (as in the case of a state or district that uses test scores for accountability purposes) scores tend to rise independent of changes in the underlying knowledge and abilities of the students (Koretz, et al., 1991). In this way, test scores can be reasonably valid measures of achievement under some circumstances but they can be corrupted when used for another purpose. Indicators built on test scores or other corruptible measures suffer a similar fate.[13]

Another concern is that variables can change meaning over time and across contexts. As noted above, as the content of courses changes, course completion means different things. As the nature of teacher preparation changes and more people enter teaching after work experiences in other fields, the college training of new teachers tells less about their qualifications. Static indicators are susceptible to

[12]Assuming the test is constructed reasonably well, administered properly and taken seriously by students.

[13]Other measures that may be susceptible to corruption include student self-reports of course-taking, teacher self-reports of inservice activities, and teacher self-reports of instruction consistent with current reform efforts.

gradual changes in their meaning, and if they are not subjected to periodic reevaluation such changes may go unnoticed.

Some threats to validity are explicitly related to indicator systems rather than individual indicators. One threat to the validity of inferences about changes in educational phenomena is that the indicator systems will be insufficiently broad to measure full range of policy-relevant concerns. This is particularly true in the case of a patchwork indicator system which relies on data gathered for other purposes. The present analysis provides a good example of this problem. Graduation requirements in mathematics and science were increased during the 1980s in an attempt to improve the preparation of students for college. One consequence of this appears to be an increase in enrollment in intermediate mathematics and science courses, such as algebra, geometry and chemistry. However, this fact alone presents an incomplete picture of the effects of this policy in a number of ways. The patchwork indicator system lacks relevant information about other concomitant changes, including changes in the content and quality of the courses, the ability of the system to meet increased demand for teachers in these areas, and the loss to these students from substituting mathematics and science for other courses, without which information about the increased course-taking could be seriously misleading.

A related threat to the validity of inferences is that indicator systems, particularly patchworks, lack data to test alternative explanations. For example, the narrowing of the gap in average mathematics achievement between majority and minority students may be due to broad advances by minority students at all levels of ability achievement, remedial efforts to raise the scores of the lowest achieving, changes in family conditions (such as family size or the educational attainment of mothers), or increases in course offerings in advanced courses in high-minority schools. An indicator system cannot in itself confirm which of these (or other) factors is responsible, but if sufficiently broad and well-constructed, it could provide suggestions about which explanations are most worth further exploration.

All of these threats to validity must be taken seriously by those considering the development of a national indicator system. In the next

chapter we consider options for indicator system development that would reduce some threats and permit researchers monitor the effect of others. One key feature is the provision of secondary sources of information to test quality and meaning of information supplied from single sources.

3. ISSUES IN DEVELOPING MATHEMATICS AND SCIENCE INDICATORS

The strengths and weaknesses of the current indicator patchwork suggest a set of principles for designing an improved indicator system. This chapter presents principles for indicator system development in the ideal case and then considers the theoretical and practical constraints that must be accommodated in building an actual indicator system. Finally, we recommend federal actions to promote the development of better indicators of mathematics and science education.

PRINCIPLES FOR DESIGNING AN IDEAL INDICATOR SYSTEM

The current indicator array - it cannot properly be called a system - reflects the diverse and sometimes conflicting purposes and principles that guided the creation of the various databases it comprises. However, if one were to design an indicator system from scratch, one would begin with a set of principles to guide the entire effort and to help mediate conflicts between competing priorities. Such a logical approach begins with broad goals and proceeds incrementally to identify constructs relevant to the goals, derive variables to embody the constructs, select measures to operationalize the variables, and so forth. The following sections describe general principles to guide indicator system development from purposes through measures, data collection and reporting.

Clarity Of Purpose

The essential first step in building a more adequate indicator system is obtaining greater agreement about purposes, because different purposes demand different information and different decisions about the fundamental design of data collection efforts. One reason currently available data do not well support a comprehensive indicator system is that they were not collected with a single purpose in mind; rather, they reflect distinct research and policy agendas. For example, SIMS and SISS (and the forthcoming TIMSS) are discipline-inspired research, while NAEP was designed as a very general education policy tool. Even subject-specific indicator efforts in mathematics and science remain

limited in scope, because they rely on common data that are being collected for other purposes, such as resource allocation, auditing, or compliance at the state level (Blank and Gruebel 1993).

Although greater clarity of purpose is a critical first ingredient in the development of an improved indicator system, complete agreement about purposes is unlikely to be attained. Even within the context of an integrated indicator system, information will be used for a variety of purposes, and these may point to different decisions about content, emphasis, or design. However, it should be feasible to develop a reasonable degree of agreement about the most important functions of an indicator system and about some priorities that can guide decisionmakers in reconciling competing goals.

We start with the premise that the fundamental function of an indicator system is description. Description is both the most important function of an indicator system and the one for which it is most adequate. Description includes static portrayals of the status of various components of education at particular points in time as well as dynamic information about changes over time in the condition of education.

Focusing on description does not restrict the indicator system to providing simple and unsophisticated information. On the contrary, the descriptive information yielded by an indicator system can be quite complex. For example, although one of the fundamental purposes of NAEP is to provide estimates of the average performance of students nationwide in three grades, NAEP can also provide estimates disaggregated by one variable at a time - for example, averages for racial/ethnic groups or for regions. To a limited degree, it also can provide multivariate descriptive information - that is, descriptive information disaggregated by more than one variable at a time, such as estimates separately by race with region or state - and it could provide more complex information of this sort as well if its sampling were altered for that purpose. In addition, NAEP's outcome variables can themselves be used to disaggregate - for example, NAEP can provide information on differences in the characteristics of teachers instructing high- and low-scoring students. All of this information,

however, regardless of its complexity, is descriptive, rather than explanatory or causal.

Within bounds, indicator data also can serve evaluative or other explanatory functions, but they are limited in this regard, and most explanatory functions should receive a relatively low priority in designing indicator systems.[14] For example, indicator data may suggest explanations that can be tested better by other sorts of information. Indicator data can also be used in conjunction with other types of information to evaluate explanations. Koretz (1987) used indicator data such as NAEP and trend data from the SAT, in conjunction with a wide array of other types of research, to evaluate common explanations of the achievement decline of the 1960s and 1970s. Finally, in certain cases, indicator data may be sufficient to evaluate explanations - in particular, to disprove them. For example, a number of common explanations of trends in the achievement of American students have assumed a specific timing of the trends, and if indicator data show that the timing was substantially different than assumed, the explanations lose credibility (see, e.g., Koretz, 1987). In designing an indicator system, however, it is important to recognize the limitations of indicator data for testing explanations and the difficulties inherent in putting them to this use.

However, even if the primarily descriptive function of an indicator system is accepted, conflicts among goals and functions will arise. Different types of description call for different designs: different constructs, variables, survey designs, etc. For example, sample size is a major constraint in most large-scale data collection efforts, because it has direct implications for cost. Therefore, a key decision is how to allocate samples to maximize the quality of the resulting information, and a decision that improves the quality of one datum can degrade the quality of others. This can be seen quite clearly in decisions that have been made with respect to the design of NAEP.

[14] This limitation may not apply to specific databases within a system. It may be practical to design databases - for example, longitudinal surveys - to serve explanatory functions and still incorporate them productively into an indicator system.

Currently, considerable resources are allocated to support the Trial State Assessment, which uses large enough samples to provide reliable estimates of performance for a limited number of assessments at the level of individual states. At the same time, the samples used to estimate long-term trends in the performance of racial/ethnic groups are small enough that many of those estimates are highly error-prone. An alternative design could reallocate resources to provide better trend estimates for racial/ethnic groups, but at the cost of the quality (or frequency) of state-representative estimates. Similarly, choices have to be made in deciding how to sample students within schools for NAEP. Currently, students are sampled randomly within schools, without regard to track, class assignments, or within-school sample size. This is a sensible strategy if one is attempting to maximize the efficiency of some population statistics, such as a national average for all students or for all students in a given racial/ethnic group. However this decision hinders the collection of information about student opportunity to learn, which is a function of class and track, and it impedes certain analyses of school-level characteristics. Initial agreement about purposes and priorities among goals will help to resolve the inherent conflicts that will arise in designing an indicator system.

Components Comprehensive with Respect to Purposes

A second principle to guide indicator system development is to include as many important constructs relevant to the purposes as feasible. Although it will often not be possible to include an exhaustive set of measures in an indicator system, it is important to include a reasonable comprehensive subset tailored to the most important purposes to which the data will be put. Here again, it is necessary to have a clear notion of the major purposes of the system and of their relative importance.

For example, an effective description of the status of mathematics and science education requires breadth. One wants to depict all the components that characterize the policy relevant aspects of the system (current as well as future): the raw materials, such as students, teachers and facilities; the action elements, such as instruction,

counseling, curriculum; and the results, in terms of knowledge and skills, states of mind, goals and expectations, etc. In comparison, a valid explanation of the relationship between curriculum content and achievement requires depth. One needs to know in detail about exposure, sequencing, multi-disciplinary overlap, opportunity to learn, instructional demands, student knowledge, student performance in a wide range of modes, etc. Failure to include relevant constructs (e.g., student characteristics, instructional quality, science exploratory skills, etc.) would severely limit the explanatory power of the indicator system. Relevant constructs include those that relate to prevailing explanations as well as those that derive from rival or alternative hypotheses.

In this regard, the development of an indicator system also is hampered our lack of a sound conceptual model of the educational system (Shavelson, et al., 1987). There is general agreement that indicators should reflect important elements of the educational system, however, without a clear understanding of how the system functions, there is no consensus on which elements are most important. In fact, importance may be derived from many things: relevance to goals, familiar ways of looking at education, proven explanatory power, relationships to other social phenomena or social programs, relevance to current policies, or an enduring role in educational policy. Each of these approaches to identifying key educational components has advantages and disadvantages, and our incomplete understanding of the underlying educational system makes it difficult to choose among them.

At present different groups are using different criteria as the basis for developing indicators. A political consensus produced the National Educational Goals. For the governors and the President who crafted the Goals the core features of the system were: readiness for school (goal 1), the school environment (goal 6), achievement in English, mathematics, science, history, and geography (goal 3, goal 4), graduation from school (goal 2), preparation for citizenship and productive employment (goal 3), and adult literacy (goal 5). These have become the central elements on which the national educational goals report is based. Although some have criticized these goals—because they

ignore important features of schooling, including the quality of educational programs, the resources available to schools and students, and students' accomplishments in art, music, drama, dance, and sports—they are not without merit. The fact that the national goals represent a political consensus about the important conditions of the nation's schools, gives them some credibility and insures some enthusiasm for their use.

There are other approaches to developing a framework for an indicator system. For example, Shavelson, et al. (1987) divided the educational system into three broad functional components: inputs, processes and outputs, indicated the functional relationship between components, and identified essential elements of each. For example, essential inputs include students, facilities, and staff; core processes include administration, curriculum, student tracking policies, and instruction. Similarly, essential outputs might include achievement, preparation for citizenship, graduation, etc. Other researchers have approached the task in a bottom up manner, proffering definitions for specific indicators of important constructs in mathematics and science education (Murnane and Raizen, 1988). Our point is that the system must be comprehensive with respect to its purposes if it is achieve any of them.

Variables Sufficient to Span Components and Permit Desired Inferences

Just as components should be comprehensive with respect to purposes, so should variables be selected to fully define components. One or more measures must be selected to instantiate each abstract component. The number of variables depends on the breadth of the construct and the inter-relationships among the variables (i.e., the degree to which they provide independent rather than redundant information about the construct). It is critical that enough relevant measures are provided to permit valid inferences about each of the components of the system.

It is impossible to say how many variables are required to provide a valid operational definition for each component. For example, even a construct as simple as attendance may demand multiple measures. Average

daily attendance (ADA) in the fall of the year is the most common attendance measure, but it does not fully portray the presence of students in schools. First, ADA changes during the school year, so it might be important to measure ADA at multiple points in time to describe trends. Second, ADA is an absolute index that does not portray the relative engagement of students. The proportion of total enrollment in attendance reflects a different aspect of attendance. In combination ADA and proportion of enrollment yield a more robust indicator of students' presence in school.

For more complex constructs the need for multiple measures is greater. For example, teacher qualifications is a high level construct that subsumes a number of narrower components, including college preparation, certification status, discipline-based knowledge, classroom experience, inservice training, and others. Each of these in turn can be measured in multiple ways. For example, college preparation can be measured in terms of degrees, major course of study, courses taken, grade point average, demonstrated knowledge on course examinations, etc. An ideal system would want to expand the construct teacher qualifications into components and include an adequate number of measures for each component.

Level of Detail Adequate to Describe Significant Differences

To be helpful for policymaking an indicator system must describe constructs in fine enough detail to reveal differences that are significant to policy. Indicators that are defined in overly broad terms, such as "years of mathematics," "science achievement," or "inservice training" are unable to reveal differences that have practically significance. NAEP provides a good example of the problems associated with level of detail. NAEP was designed to provide general information about student achievement in mathematics and science, but not to differentiate at the level of topics or subtopics. The deficiency is the greatest in terms of science; students receive an overall score in science, but the assessment does not provide scores in biology, chemistry, earth science, etc. Since much science instruction

is structured along these disciplinary lines, lack of information on achievement represents a severe limitation.

An indicator system requires enough detail to distinguish differences that are meaningful in practice. This requires that designers be clear about the purposes to be served by the indicator system and understand the phenomena well enough to know which construct-related differences are important. For example, a survey designer might think it a good thing to operationally define curriculum in terms of years of coursework; it is an efficient way to aggregate data. However, a math or science educator would understand that two years of basic mathematics is not the same as two years of advanced mathematics, and an indicator that equates the two provides an incomplete, and even misleading, picture of exposure to curriculum.

Data Collection Strategies and Formats That Assess Full Range of Constructs

In operationalizing the variables to be measured, it is important to use data collection strategies that assess the construct in the most complete and appropriate manner feasible, given resource constraints. For example, until recently NAEP used multiple choice items to assess science and mathematics achievement. Well-designed multiple choice questions can measure some elements of higher-order thinking, such as the ability to integrate, evaluate and synthesize information, as well as factual knowledge for which they are commonly used. However, they measure these ability based on respondents selections from a fixed set of choices. This is a limited range of cognition. It might be more effective to measure directly a person's skill at performing mathematical and science tasks, such as conducting an experiment, solving a problem, manipulating apparatus, finding an error in a proof, etc. The current interest in "alternative" or performance based assessment stems in part from the wide recognition that different forms of testing are appropriate for measuring different intellectual abilities. This is not new knowledge (citation) but it is only recently being used to overturn years of dominance based on the practical advantages of multiple choice tests.

Most constructs can be measured in multiple ways, and an ideal indicator system will use strategies that give the fullest (and hence most valid) reading of the underlying construct. For example, McDonnell et al. are investigating ways to use classroom artifacts (tests, assignments, logs) to measure students exposure to specific mathematics and science principles. At present curriculum information is gathered through surveys that ask teachers whether students had opportunities to learn specific problems in class (SIMS) or whether specific topics were presented as a major topic, a minor topic, a review topic or not covered during the course (Horn, Hafner, and Owings, 1992). These alternative formats may reveal more about the opportunities students have to learn specific content than surveys.

The general point is that measurement should utilize strategies that reflect the construct as fully as possible. These approaches may not be the most economical—current limited use of performance assessments to measure student achievement cost more than multiple choice tests (General Accounting Office, 1993) and wider use could cost many times more (Koretz, et al., 1992; Madaus and Kellaghan, 1991; Stecher, 1995)—but they operationalize the construct in more meaningful ways and provide information that may be more valid.

Data Collection That Permits Aggregation at Appropriate Levels and for Appropriate Subgroups Vis-A-Vis Purposes

An ideal indicator system would sample extensively enough to permit valid summaries at multiple levels of analysis and for multiple policy-relevant subgroups of students. National averages seldom provide the information policymakers or researchers need to address educational issues. Even the decline in SAT scores, which is frequently portrayed at a national phenomenon, is unrevealing until it is disaggregated by the characteristics of students in the test taking pool (Koretz, 1987, 1992b). In policy terms, it might be said that the "average" case provides a general thermometer, but the special cases are the diagnostic and prescriptive tools. Furthermore, relationships at one level of aggregation can differ from those at another level of aggregation, therefore it is important to be able to look at multiple levels (Langbein and Lichtman, 1978). Consequently, data collection to support

an indicator system must be sensitive both to levels of analysis and subgroups of interest.

However, the ability to report valid information at lower levels of aggregation or for population subgroups has a price; analyses of smaller units requires larger overall samples. It also entails trade-offs: disaggregating on any one dimension consumes sampling and other resources that could have otherwise been used to disaggregate on another. For example, because of the large increase in resources required to provide state-representative estimates, designs that provide them are likely to reduce either the scope or the quality of national statistics.

The value of national and state reporting is great because educational policymaking is primarily a state and national responsibility. In fact, states are responsible for the lion's share of educational resources and control the bulk of the regulatory apparatus—graduation requirements, teacher certification standards, facility standards, etc.--and it would be uninformative if each state's status could not be reported. As a result, an ideal indicator system would permit aggregation of many results at both the state and national levels. In contrast, geographic regions traditionally have little meaning as policy units, and regional reporting seems unnecessary. Where states have banded together into regional consortia and undertaken coordinated policies, aggregated information may be meaningful (Bottoms, et al., 1992). However, arbitrary regional aggregations seem superfluous; there has been little interest over the years in NAEP's regional data summaries.

The case for district-level reporting is not a clear; it runs afoul of practical constraints. Districts are vested with considerable decision making authority in most states, so it might be desirable to have district-level reporting, as well. However, widespread reporting at the district level is impractical because the costs of such an extensive data collection effort would be prohibitive. One might envision special supplemental studies in which a sample of districts were included, but not as a core element in an indicator system.

It is also important that data collection be planned with an eye towards the population subgroups of policy interest. This includes racial/ethnic groups, as well as subpopulations that have important roles in terms of the mathematics and science education system, e.g., high achieving students, students who take advanced mathematics and science courses, teachers who meet national association standards for certification in mathematics and science, teachers who enter the profession through alternate routes, etc. To draw inferences about such subgroups decisions have to be made at the sampling stage to ensure adequate representation. As in the case of state or district disaggregation, each such decision has cost implications.

Frequency of Data Collection Inversely Proportional to Rate of Change in Construct

In an ideal indicator system, there would be an inverse relationship between the frequency of data collection and the degree of fluctuation in the measure. Just as scientists measure the movement of a glacier less often than they measure the movement of the tides, an indicator system can adjust the frequency of data collection to reflect the rate of change of the constructs being measured. The recognition that it is not necessary to gather data annually on all constructs and all subgroups can help to reduce costs. For example, one might collect data on enrollment and teacher characteristics biennially, information about graduation requirements every four years, but only conduct detailed studies of selected course content every six to eight years.

However, the goal of matching the frequency of data collection to the rapidity of change in the underlying constructs may difficult to achieve in practice. While historical trends can be used to establish expectations regarding the volatility of many measures, unanticipated changes can occur to bedevil indicator planning. This is particular true for constructs that reflect policy decisions, e.g., requirements, standards, and capacities. For example, regular graduation requirements in mathematics and science which increased rapidly between 1980 and 1985, have remained relatively stable thereafter (Stecher, 1992). Similarly, the publication of *A National at Risk* and other critical analyses of the U.S. educational system sparked two years of widespread

state-level educational reform (Coley and Goertz, 1990). Nothing in the previous few years would have predicted these rapid changes. Such unanticipated fluctuations have greatest policy implications, but they are unpredictable.

Changes in "human" features of the system, such as student demographic characteristics, teacher supply, etc. tend to be gradual, and are less likely to be affected by short-term policy initiatives. In this domain the speed of change is more predictable, so it should be possible to accommodate rates of change in the design, i.e., to conduct infrequent measurement of relatively constant factors and more frequent assessment of more volatile features.

Measurement Quality Adequate for Inferences (Multiple Independent Sources of Data)

An ideal indicator system will provide data whose technical quality is adequate to support the inferences for which the system was designed. There are two ways in which quality must be adequate—the system must present data which contain a minimum of error due to measurement, sampling, or other factors, and the system must provide multiple independent sources of information to permit users to assess the accuracy of data sources and the validity of inferences. These quality concerns reflect the traditional notions of reliability and validity as applied in the context of indicators.

An ideal system requires data that are highly reliable, a goal which is not always achieved in practice. Our patchwork analysis revealed a number of weaknesses in the core data (Koretz, 1991; Stecher, 1992). For example, some measurement techniques, such as self-reports, are commonly used in large-scale data collection despite their proven susceptibility to errors. Comparisons between transcripts and self-reported course taking reveal a pattern of errors ranging from minimal to potentially significant (Valiga, 1986; NCES, 1984). An ideal indicator system also requires data that are representative of the populations of interest. Although most national surveys are designed to be representative, some relevant data collection efforts are not. For example, the IEA Mathematics and Science studies of the 1980s were not conducted on representative samples, which severely limited their value

for indicator construction. Even surveys designed to be representative may prove problematic if response rates are low or data are incomplete. We found enough missing background data on NAEP school reports to question the representativeness of the findings. As noted earlier, surveys also must be representative of policy-relevant subgroups. Unfortunately, measurement and sampling errors are not always apparent to the users of the data; it can require considerable effort as well as sophistication to uncover such problems. Discrepancies between independent sources of information are often the first clues that errors of measurement exist, so multiple sources become an important way of insuring the quality of indicator data.

An ideal indicator system would include multiple independent measures of important constructs, both to test the quality of data and assess threats to the validity of interpretations drawn from indicators. It is only through comparisons between data sources that problems with data collection procedures, operational definitions and other quality concerns are revealed. More importantly, multiple sources of information are essential for establishing the validity of inferences from indicators. The debate about trends in the achievement of students in the U.S. is a perfect example of the difficulty of relying on single sources of data. Even a simplified comparison of two data streams—standardized test scores and college entrance examination scores—reveals some of the problems inherent drawing inferences from indicators. Standardized test scores, such at the Iowa Test of Basic Skills, portray different trends than SAT scores. The differences derive from a number of factors including underlying differences in the variables measured, incomparability of the samples, and corruptibility of standardized test score in high stakes testing programs. Relying on one source alone would fail to reveal these problems. In most cases, it is impossible to determine whether inferences are justified without ways to test for robustness across operational definitions, samples, and conditions. An ideal indicator system would include some level of redundancy, either in the form of parallel studies or periodic supplementary data collection focused on the most important variables and those thought to be the least robust.

Measures Sensitive to Changes in Phenomena Under Study

This principle seems rather simple—monitor things in a way that will detect changes; yet it can be a difficult principle to follow because of shortcomings in the way we define and measure features of the educational system. Many of the features for which we may want to build indicators are defined in discrete ways that may be insensitive to potentially important changes in the underlying phenomenon. A teacher's credential status is a dichotomous variable (that changes only upon receipt of a credential) although the teacher's formal preparation grows incrementally from course to course and workshop to workshop. Measures of teacher preparation defined in terms of credential status are insensitive to staff development activities that may significantly enhance teacher quality. The problem is less severe for features that are measured in continuous ways, but even measures that are continuous are not equally sensitive to changes at all levels. For example, the NAEP achievement tests contain fewer items of high and low difficulty than of moderate difficulty. As a result, we are less able to detect changes in the performance of "high achieving" students.

Even with good intentions, we may strain the sensitivity of measures when we examine fine-grained questions about subtopics and subgroups. Indicators of student achievement provide good examples of this problem. Koretz (1991) describes the interactions between population sampling (to provide subgroup scores), content sampling (to provide topic scores), and difficulty level (to differentiate at relevant points in the distribution). Differences between subject area composites and specific content areas depend both on the groups for which contrasts are drawn and the subject under investigation. An ideal indicator system would be defined in terms of measures that were continuous, covered the full range of policy-relevant values, and were equally sensitive to changes throughout that range.

Analyses and Presentations That Reveal Underlying Relationships

An ideal indicator system should strive to analyze and display data in ways that maximize the clarity, comprehensiveness, and utility of the information (Koretz, 1991). Published indicator data typically describe

performance at the mean, but such measures of central tendency are not always sufficiently revealing. In many cases it is necessary to compare additional aspects of the distributions.

The domain of student achievement provides a wealth of examples of the value of alternative presentations. For example, differences in central tendency between groups can be put in comparative forms that provide both an intuitively clear indication of the magnitude of the differences and a small amount of information about the degree of overlap between distributions (Koretz, 1991). One such option is the proportion of students in one group who score above the median of a second. Another representational indicator that can clarify the information for lay audiences is the percent of individuals from different groups scoring in the top N% overall. Many of the observers using indicator data will not anticipate the progressively sever underrepresentation of low-scoring groups (although this is predictable from the underlying distributions). The routine use of such comparative metrics would appear to be desirable for indicator systems, for it increases the clarity of the information presented without increasing the complexity or complicating the display of trends.

Supplemental presentations that reveal more about the underlying distribution of data, such as box plots, may be necessary when these distributions are unusual, e.g., when distributions are not symmetrical or differ in more than central tendency, such as within group variances. Graphical displays of entire distributions of performance may be appropriate for certain situations, but they are complex and this complexity limits their utility, e.g., they cannot easily be used for monitoring trends.

Some decisions about how best to display data may depend on the uses to which specific indicators are put. For example, simple subject-area composite scores are generally sufficient for school-level indicators, but their adequacy for the more common student-level indicators varies depending on the subject and the groups contrasted (Koretz, 1991). The ideal indicator system would adapt displays to portray the underlying information in the clearest and most revealing manner. The tradeoff between simplicity and comprehensiveness will vary

from one instance to another, depending on the aspects of performance that are being measured and the groups for which performance is being contrasted.

CONSTRAINTS ON INDICATOR SYSTEMS

The previous discussion seems to suggest that an indicator system should encompasses a large, diverse set of constructs each measured by multiple variables at fine levels of detail, with overlapping, independent data collection efforts operating in parallel. However, there are theoretical reasons why this approach is not optimum and practical limitations that make this goal unattainable. Both sets of constraints need to be understood before making decisions about national indicators for mathematics and science education.

Theoretical Limitations

There is fundamental tension between simplicity and comprehensiveness that is inherent in the definition of indicators. By design, indicators are simple statistics, but they are valued as a way to understand diverse, complex systems. An immediate challenge in developing indicator systems is to balance simplicity and comprehensiveness. A desire for completeness and explanatory power argues for increasing the number of variables that are included, the number of ways each is measured, and the level of detail of observations. However, indicator systems are valuable because they are limited, succinct and parsimonious. The purpose of indicators is to illuminate key elements of larger phenomena in a simple and concise manner, and this purpose precludes measuring comprehensively. One cannot achieve both goals; compromise is required.

Policymakers appear to value parsimony. For example, the National Educational Goals Report focuses on only six goals, and for each goal it presents only a handful of measures. Efforts to add additional goals were defeated in the name of simplicity and consensus. Similarly, in an effort make reporting more relevant to policy, NAEP reports data in

terms of three achievement levels.[15] In these cases and others it is clear that policymakers preferences lean toward simplicity, and indicators must meet policymakers needs if they are to receive their political and financial support (Shavelson, 1987).

Practical Constraints

In addition, large, comprehensive indicator systems are expensive, and resources for their development are limited. Shavelson, et al. (1987) estimated the cost for a comprehensive independent indicator system to be between $23 million and $34 million. There is no indication that the National Science Foundation or the US. Department of Education is likely to fund an effort of this size at present, nor would they have committed this level of resources to indicators at any time during the past decade. These fiscal limitations translate into fewer variables and more limited measurement strategies.

The demands of implementing and managing an indicator system also create procedural limitations on its design and scope. These factors manifest themselves in a number of ways. For example, too much measurement is an annoyance to schools, and, educational administrators already are chaffing at requests for more data collection. Complex approaches to data collection, such as the use of instructor logs or classroom archives, must often be tempered by the practical constraints of time, training, ownership, standardization, and commitment. In addition, indicator systems must be managed in light of changing purposes and evolving features of the educational system, and resources must be devoted to this ongoing management function. Designers should not conceive of an indicator system as a static entity, put into motion and left to operate with little intervention. I

The challenge to indicator system designers is to craft the most effective compromise given these theoretical and practical considerations. There is no simple resolution to the tension between the goals of simplicity and comprehensiveness; neither is there a way to escape the constraints imposed by budgets and administrative concerns.

[15]Despite objections from OTA and the NAEP Technical Review Panel that the levels are invalid.

In the following section we offer suggestions for developing and managing an improved indicator system for mathematics and science education that aspires to the ideal while recognizing the realities delineated above.

RECOMMENDATIONS TO NSF FOR IMPROVING INDICATORS OF MATHEMATICS AND SCIENCE EDUCATION

In a previous study, RAND described five indicator system options; here we recommend that NSF adopt a hybrid approach that combines features from two of those options. The five options were called status quo, patchwork, cyclical studies, piggyback and independent, and they spanned a continuum from very low NSF involvement to comprehensive intervention (Shavelson, et al., 1987). The status quo option represented the lowest level of commitment; NSF would use whichever data were available when a policy question arose. The independent option was the most comprehensive, calling on NSF to develop an independent data collection and analysis system of its own. The authors of that report dismissed the status quo option because it was not an indicator system at all, and they argued that the independent option was too expensive to be practicable (Shavelson, et al., 1987). They suggested that NSF give serious consideration to the three options in the middle.

On the basis of the present study we recommend that NSF adopt a "supplementary" approach that combines some of the elements of the cyclical studies and piggyback approaches described previously. Specifically, we suggest that NSF use its resources to supplement existing data collection efforts to obtain more complete data in areas of interest (piggyback) and commission additional data collection on a periodic basis to provide longitudinal measures, secondary sources, and in-depth data that are not available through large-scale efforts (cyclical studies). This combination of approaches offers the best chance of creating indicators that, taken as a whole, approach the functionality of a comprehensive indicator system.

Our view of NSF's role reflects assumptions about NSF's responsibilities for indicators and our view that the current indicator system is deficient and in need of research and development. We assume that NSF's role is more circumscribed than that of statistical agencies

such as the National Center for Education Statistics (NCES). Of course, NSF is primarily interested only in mathematics and science education. Beyond that, NSF does not have responsibility for routine and cyclical data collection (such as the Common Core of Data maintained by NCES). The narrower purview of NSF provides an opportunity; lacking certain routine data collection responsibilities and free to concentrate on a few subject areas in depth, NSF has the prerogative to be more forward-looking in its approach to indicators.

Given these premises, we urge NSF to be the most forward-looking agency in the federal education indicator effort. NSF should assume central responsibility for managing the indicator effort and coordinating key functions primarily performed by others, including design, data collection, research and reporting.[16] It should leave to others most of the operational responsibilities for design, data collection, analysis, and reporting of routine data collection efforts. This role is essential if a supplementary system is to function effectively, because such a system must be built from disparate parts, and this can occur only if someone is actively managing the effort.

NSF has both the interest and the experience necessary to perform this function. In fact, most mathematics and science indicator research and development has been sponsored by NSF, as have many of the large-scale data collection efforts that provide information on which indicators can be built. As the central point of contact, NSF can voice the inherent tradeoffs that must be addressed in indicator development, negotiate compromises, and maintain the larger perspective that is sometimes lost by those who are engaged in data collection, analysis and interpretation.

We believe that NSF should focus much of its indicator efforts in two ways. First, it should support diverse supplementary data

[16]Apologies for using an athletic metaphor, but it was the most apt comparison we could produce. Another image that comes to mind is the illustration from the musical comedy *My Fair Lady;* it shows a hierarchy of three marionettes, with George Bernard Shaw holding the strings that support Henry Higgins, who in turn holds the strings that support Eliza Doolittle. It doesn't work quite as well, but it is free of competitive connotations.

collection efforts, including add-ons to routine data collection and additional special studies. This is a traditional role for NSF reflected, for example, in its support for such efforts as TIMSS. Second, NSF should support planning, research and development, and evaluation pertaining to indicators. This latter focus would include, for example, experimental uses of new indicators, validation research, and periodic benchmarking studies.

Develop an Infrastructure to Support and Improve Indicators

We also recommend that NSF develop an advisory infrastructure to guide its actions vis-à-vis indicator design and development. To this end, NSF should create a standing Indicator Advisory Group (IAG) with responsibility for monitoring supplementary indicator efforts. The IAG should undertake tasks such as building consensus about purposes, evaluating existing data collection activities, establishing priorities for supplemental data collection, communicating with other agencies to increase the utility of their efforts, conceptualizing new studies that would address issues not covered by available efforts or test the reliability and validity of existing data, and monitoring indicator-related efforts at the national level. The group should view its key responsibility as diagnosis and improvement, i.e., asking critical questions, identifying shortcomings, gaps and problems, and recommending actions to resolve them. Members should include researchers and representatives of relevant federal agencies and research organizations. We also suggest that the standing IAG be supplemented as appropriate with *ad hoc* committees with specific foci or expertise.

The importance of an independent coordinating body such as the IAG is clear from a review of indicator data from the past decade. Indicator-like reports have been the products of episodic secondary analyses of data collected for other purposes, and as a result they have been incomplete and of limited value. The *National Educational Goals Report* (NEGP, 1993), arguably the country's most visible indicator effort, is a hit and miss collection of information gathered for other purposes, and it's existence reflects a tenuous political consensus whose life is unpredictable. NCES has published educational statistics

for years in *The Digest of Educational Statistics* (NCES, 1994), but only recently have they begun to select and focus the information in the manner of indicators. Even so, these statistics reveal little about mathematics and science education. *Science and Engineering Indicators* (NSB, 1993), which may be the oldest educational indicator publication, is a more stable effort to collect what is available and mold it into an interpretable pattern. Only the Council of Chief State School Officers *State Indicators of Science and Mathematics* (Blank and Gruebel, 1993) demonstrates conceptual planning and coordination, although it's scope is limited to state-level data. Nevertheless, this effort demonstrates the value of more focused planning, design and monitoring. These are the functions on which NSF should concentrate if it wants to promote more useful mathematics and science education indicators.

It is beyond the scope of this paper to design a structure and define specific responsibilities for the IAG, but we believe its initial efforts should include clarifying the purposes to be served by the indicator system and developing priorities and operating principles for NSF supplemental indicator efforts. As we pointed out previously, policymakers hope to use indicators for a variety of different purposes, which dictate different strategies for indicator system development. For example, the relative emphasis that is placed on breadth versus depth or measuring constructs at extreme values versus measuring them at more typical values depends on the questions to be answered and the uses to which those answers will be put. Consequently, it is important to establish priorities among purposes before deciding on actions to improve mathematics and science indicators.

We believe that description is the most appropriate purpose for mathematics and science indicators, and that NSF should focus its efforts on this goal. Past experience suggests that inflated expectations for indicators are unlikely to be met (Shavelson, 1987); an indicator system will not provide valid explanations of educational changes nor will it admit inferences about causes and effects. However, with support and guidance from NSF it should be possible to develop an effective indicator system to monitor changes in key features of the educational system over time. This perspective argues that NSF should

focus its effort on promoting regular, consistent, and relevant descriptive data.

However, the decision about which purposes to serve should not be made without some consideration of policymakers' needs. It would be unwise to exclude the policymakers' perspective from deliberations about indicator development. This does not imply that the policymakers' goals will carry the debate, but that an attempt should be made to achieve a consensus on purposes based on inputs from all interested parties. This effort will increase the utility of whichever indicators ultimately are produced.

Unfortunately, a consensus on purposes for mathematics and science indicators will not be easy to achieve. McDonnell's (1993) description of the complex motives that underlie assessment policy presents a pessimistic picture. McDonnell argues that informational policy tools, like an indicator system, receive broad support precisely because they serve multiple purposes. Policy makers agree on specific assessment policies while holding very different goals for the data. Consequently, while it is important to have purposes in mind to guide indicator system development, it is unlikely that a clear consensus on purposes can be achieved.

Nevertheless, it should be possible for the IAG to set priorities among goals to guide planning for indicators, while admitting that users' needs differ. Indicators will be used for multiple purposes regardless of the intentions of indicator developers. Establishing priorities among goals offers some basis for practical decisions such as the choice of constructs and variables and the development of sampling strategies. Priority setting may be a contentious process, but even a compromise result that admitted multiple uses while assigning greater priority to some would be a better guide to planning than no resolution at all.

A second key tasks to be performed by the Indicator Advisory Group is to establish operating principles to serve as a basis for decisions and recommendations to NSF. For example, one of the principles that appears useful based on our patchwork study is to assign higher priority to validation studies and research that will not be done by others. The

IAG must decide if it endorses this focus on research to supplement, test and validate the existing indicators. A second principle might be to emphasize data collection and analyses that fill in holes in the existing patchwork. For example, NSF might fund occasional data collection supplements to over-sample specific populations and sub-populations and permit more in-depth analyses. The IAG also might want to give priority to developing and testing potential new indicators. A third operating principle might be that NSF should continue to rely on other organizations to carry out the fundamental research activities necessary to build indicators, including sampling, data collection, analysis and reporting. Universities, federally-funded centers and labs, research organizations and other federal agencies execute the strategy with NSF occasionally pitching in to help. By engaging a wide range of organizations, NSF maximizes the talents brought to bear on indicator issues. As a secondary benefit, this approach preserves the capabilities that exist in other organizations contributing to the overall research capacity and diversity.

In setting policies and recommending supplemental research, the IAG will have to address the relative cost and value of alternative indicators and research activities. Some indicators that are important are relatively inexpensive to collect, such as graduation requirements, while others that may seem less important upon first glance are extremely expensive, such as within-student growth in science achievement. Although it may be possible to put a price on each variable in terms of data collection and analysis, it is unlikely that the value of each can be easily reckoned. The advisory committee will have to attempt balance costs against potential policy import. In each case there probably is no single optimum solution, and the best strategy is one that maintains a reasonable long-term balance between competing objectives.

Because the supplemental indicator approach permits a long-term view, the IAG can address issues that are unlikely to be resolved under the present system. For example, one of the great shortcomings of federal data collection is its reliance on cross-sectional data. This occurs because the cost of true longitudinal research at the student or

teacher level is quite high. Only someone with an external focus is likely to conduct research to resolve differences between surveys or to assess the relative validity of alternative operational definitions or approaches to data collection.

The Value of Consistency

Consistency of planning and funding is needed to overcome the irregularity and volatility of data that make a patchwork indicator system unstable. One of the advantages of the approach recommended here is it promotes forethought in planning so indicator related data are more likely to be available regularly and to be comparable from cycle to cycle. There also is the need for consistency in data collection so critical information is provided in timely fashion. Therefore, it is important to maintain funding for the Indicator Advisory Group and for key research activities for an extended period of time. Similarly, members of the IAG should be serve staggered, overlapping multi-year terms to increase the consistency of planning. By staying the course, NSF will enhance the value of its indicator efforts. Without continuity of funding, purpose and leadership, we will continue to have a haphazard patchwork rather than an indicator system.

A. ADDITIONAL DATA DESCRIBING THE TEACHER WORKFORCE

Table A.1

Certification Status, Public Secondary School Teachers Whose Primary or Secondary Assignment is Chemistry, Physics, or Earth Science

Subject and grade level	Percent of teachers fully certified in subject*	Percent of teachers initially certified in subject+	Percent of teachers excluded from tally**
Chemistry			
Grades 9-12	93	95	19
Physics			
Grades 9-12	94	95	70
Earth science			
Grades 7-8	81	86	36
Grades 9-12	80	87	32

*Fully certified includes teachers with "regular or standard" state certificates and teachers with probationary certificates, defined as "the initial certificate issued after satisfying all requirements except the completing of a probationary period."

+Initially certified includes temporary or emergency certification in addition to the other types.

**This column indicates the percentage of mathematics or science teachers of each type who taught the subject only occasionally and were excluded from the other tabulations

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.2

Certification Status, Public Secondary School Teachers Whose Primary or Secondary Assignment is Mathematics or Science, Grades 7-12 Combined*

	Percent fully certified in subject+		
Subject	Primary assignment field	Secondary assignment field	Either primary or secondary
Mathematics	93	67	91
Biological science	95	74	92
Physical science	89	77	85
General science	90	75	86

*Teachers who teach mathematics or science occasionally are not included

+Fully certified includes teachers with "regular or standard" state certificates and teachers with probationary certificates, defined as "the initial certificate issued after satisfying all requirements except the completing of a probationary period."

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.3

Percentage of Secondary School Teachers with Highest Attained Degree at Each Level

	Mathematics teachers		Science teachers	
Highest degree attained	Grades 7-8	Grades 9-12	Grades 7-8	Grades 9-12
Public school teachers				
Associates degree	0	0	0	0
Bachelors degree	53	48	53	42
Masters degree	41	45	39	48
Educational specialist	5	6	8	8
PhD/professional degree	1	1	1	1
Private school teachers				
Associates degree	1	1	0	0
Bachelors degree	56	48	61	50
Masters degree	37	49	34	38
Educational specialist	4	2	4	4
PhD/professional degree	2	1	1	8

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.4

Percentage of Public Secondary School Science Teachers with a Major or Minor in Specific Science Field of Study or Science Education

Major/minor (degree) area and subject taught	Grade 7-8 Subject [a]	Grade 7-8 Either	Grade 9-12 Subject [a]	Grade 9-12 Either
Degree in science				
Science teachers	50	64	73	86
Biological science teachers	53	66	79	92
Physical science teachers	51	66	74	76
General science teachers	38	53	55	72
Degree in specific science field taught				
Biological science teachers	45	60	73	88
Physical science teachers	32	51	56	73
Chemistry teachers	NA	NA	59	78
Physics teachers	NA	NA	33	56
Earth science teachers	11	36	30	52

[a] "Subject" column lists percent of teachers who have a major or minor in the degree area; "either" column lists percent who have a major or minor in the degree area or in science education. The SASS provides sample sizes large enough to examine physics and chemistry teachers only at grades 9-12.

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.5

Percentage of Public Secondary School Mathematics and Science Teachers with the Equivalent of a College Major or College Minor in Their Subject Field

	Major-equivalent indicators		
	Strict major units (20+)	Loose major units (13+)	Major or minor units (8+)
Mathematics			
Grades 7-8	17	36	62
Grades 9-12	23	51	80
Biological science			
Grades 7-8	19	32	53
Grades 9-12	28	51	77
Earth science			
Grades 7-8	6	14	28
Grades 9-12	14	26	48
Chemistry			
Grades 9-12	20	39	58
Physics			
Grades 9-12	11	26	51

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.6

Percentage of Secondary School Mathematics and Science Teachers with Recent Training Experiences, 1985

	Mathematics		Biological sciences		Physical sciences	
Training activity	7-8	9-12	7-8	9-12	7-8	9-12
Any training in 1984	59	55	59	54	53	43
Type of training*						
College course	33	34	33	36	32	31
In-service	34	30	33	30	26	21
Other	6	6	9	8	8	5
Purpose of training						
Maintain/improve abilities	48	44	51	48	42	36
Retrain	9	8	5	4	7	5
Nonteaching credentials	3	3	3	2	3	3

*Teachers could respond with more than one type of training.
SOURCE: 1985 Public School Survey.

Table A.7

Percentage of Public Secondary School Teachers Taking 30-Credit Equivalent Training in Past Two Years, for Any Assignment Field, and for Their Teaching Field

	In-service in:	
Teaching field	Any field	Teaching field
All areas	35	N/A
Mathematics	34	14
Science	38	15
English	35	13
Social studies	32	14

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.8

Percentage of Public and Private Secondary School Mathematics/Science Teachers Taking Given Number of College Courses in Mathematics/Science

Number of courses taken	Mathematics teachers		Science teachers	
	Grades 7-8	Grades 9-12	Grades 7-8	Grades 9-12
0	3	2	3	2
1-2	9	5	3	1
3-5	18	7	12	4
6-12	36	35	20	12
13-19	18	27	19	22
20+	16	23	43	59
Median	9	12	17	21
Mean	9	14	21	26

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.9

Percentage of Public Secondary School Biological, Physical, and General Science Teachers Taking Given Number of College Courses in Science

Number of science courses taken	Grade 7-8 teachers of:			Grade 9-12 teachers of:		
	Biological science	Physical science	General science	Biological science	Physical science	General science
0	1	4	5	2	2	1
1-2	1	1	0	2	2	3
3-5	10	7	17	1	2	9
6-12	21	20	18	10	11	12
13-19	19	23	17	22	22	24
20+	49	46	43	63	61	50
Median	20	19	18	23	22	19
Mean	24	22	21	27	26	25

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.10

Percentage of Public Secondary School Science Teachers Taking Given Number of Courses in Areas of Science Related to Their Assignment

Number of courses taken	Teaching field: Biological science	Physical science	Earth science	Chemistry	Physics
Grades 7-8					
0	3	4	14	NA	NA
1-2	14	9	29	NA	NA
3-5	19	19	20	NA	NA
6-12	31	35	24	NA	NA
13-19	13	13	8	NA	NA
20+	19	20	6	NA	NA
Median [a]	8 (20)	9 (18)	3 (20)	NA	NA
Mean [a]	12 (24)	12 (22)	6 (22)	NA	NA
Grades 9-12					
0	2	3	13	3	3
1-2	5	3	22	3	10
3-5	8	7	12	22	28
6-12	35	32	27	33	32
13-19	23	23	12	19	15
20+	28	33	14	20	11
Median [a]	12 (23)	14 (22)	6 (21)	10 (23)	7 (23)
Mean [a]	15 (27)	18 (26)	6 (25)	13 (29)	9 (27)

[a] Numbers in parentheses are the median and mean number of total science courses taken by teachers in each science field.

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.11

Demographic Composition of Grade 7-12 Public School Mathematics and Science Teachers Compared to All Public School Teachers

Teachers	Percent female	Percent minority	Percent non-Asian minority
Mathematics and science teachers	47	10	9
All teachers	54	8	6

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.12

Demographic Composition of Grade 7-12 Public School Mathematics and Science Teachers Compared to New Mathematics and Science Teachers

Teachers	Percent female	Percent minority
Mathematics		
All teachers	52	10
New teachers*	59	9
Sciences		
All teachers	42	9
New teachers*	52	8

*New teachers are defined as those in their first year of teaching, with no previous teaching experience.

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.13

Percentage of Public Secondary School Science Teachers Who State that They are Best or Second Best Qualified to Teach "Science"

Subject and grade	Best qualified field	Second best qualified field	Neither
Biological science			
Grades 7-8	74	12	14
Grades 9-12	83	8	9
Physical science			
Grades 7-8	69	12	19
Grades 9-12	82	8	10
General science			
Grades 7-8	62	16	22
Grades 9-12	71	17	22

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.14

Percentage of Grade 7-12 Public School Teachers Primarily Assigned to Subject Who Have Switched from Another Primary Assignment Field

Current primary assignment field	Percent with assignment change
Earth science	53
Physics	51
Chemistry	49
General/other science	41
Biology	36
Mathematics	24
English	20
Social studies	18

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.15

Previous Primary Assignment Fields of Grade 7-12 Public School Mathematics and Science Teachers who have had an Assignment Change

Previous primary assignment field	Percent of those whose current primary assignment field is:					
	Math	Biology	Chem-istry	Earth science	Physics	General science
Mathematics	N/A	5	18	7	23	8
Biology	5	N/A	29	20	14	20
Chemistry	3	7	N/A	4	15	7
Earth science	2	12	5	N/A	0	5
Physics	3	1	8	1	N/A	2
General/ other science	10	33	33	22	32	N/A
Health/physical education	7	12	1	6	0	17
English	8	7	2	6	7	3
Social studies	10	4	2	4	2	4
Vocational education	7	3	1	3	0	5
Elementary education	32	10	2	20	3	19
All other areas	14	6	0	7	4	11
Any science field	N/A	53	75	47	61	34

Source: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.16

Previous Main Activities of Newly-Hired Inexperienced Public Secondary School Mathematics and Science Teachers

	Percent engaged in each activity	
Activity	Mathematics and science teachers	All teachers
Working in nonteaching education position	3	7
Working outside of education	19	17
Teaching—other school	7	5
Teaching—this school	—	—
Postsecondary teaching	—	—
Homemaking/child care	4	5
Attending college	60	59
Military service	1	1
Unemployed/retired/other	6	6

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

Table A.17

Former Occupations of Grade 7-12 Public School Mathematics and Science Teachers

Occupation	Percent
Mathematics and science fields	21
Health, science, and other technicians	6
Architects/engineers	5
Natural scientists	4
Farming/forestry/fishing	4
Mathematicians/computer scientists	1
RNs/therapists	1
Physicians	0
Post secondary teachers	0
Other fields	79
Sales occupations	20
Managers/administrators	15
Administrative support	14
Craft and repair occupations	9
Service occupations	7
Operators/laborers	7
Social/religious workers	4
Writers/artists/athletes	3
Librarians	0
Social scientists	0
Lawyers/judges	0

SOURCE: RAND analyses of the 1987-88 Schools and Staffing Survey.

BIBLIOGRAPHY

Adleman, C. (Ed.) (1989). Signs and traces: Model indicators of college student learning in the disciplines. Washington, D.C.: U.S. Department of Education.

Barron, S. I and Koretz, D. M. (forthcoming). An Evaluation of the Robustness of the NAEP Trend Lines for Racial/Ethnic Subgroups. *Educational Assessment*.

Beaton, A. E., and Zwick, R. (1990). Disentangling the NAEP 1985-86 Reading Anomaly. Princeton: NAEP/Educational Testing Service.

Berends, M., and Koretz, D. M. (forthcoming). Measuring Racial and Ethnic Test Score Differences: Can the NAEP Account for Dissimilarities in Social Context? Santa Monica, CA: RAND.

Blank, R. K. and Dalkilic, M. (1990). State indicators of science and mathematics education, 1990. Washington, D. C.: Council of Chief State School Officers.

Blank, R. K. and Gruebel, D. (1993). State indicators of science and mathematics education, 1993. Washington, D. C.: Council of Chief State School Officers.

Bottoms, G., Presson, A., and Johnson, M. (1992). Making high schools work through integration of vocational and academic education. Atlanta: Southern Regional Education Board.

Bracey, G. W. (October, 1992). The Condition of Public Education. Phi Delta Kappan, 74,(2), 104-117.

Bracey, G. W. (October, 1993). The third Bracey report on the condition of public education. Phi Delta Kappan, 75,(2), 104-117.

Burton, N. W., and Jones, L. V. (1982). Recent trends in the achievement levels of black and white youth, Educational Researcher, 11 (April), 10-14, 17.

Cannell, J. J. (1987). Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above The National Average. Daniels, WV: Friends for Education.

Cannell, J. J. (1989). The "Lake Wobegon" Report: How Public Educators Cheat on Standardized Achievement Tests. Albuquerque, NM: Friends for Education.

Choy, S. P., Bobbitt, S. A., Henke, R. R., Medrich, E. A., Horn, L. J., and Lieberman, J. (1993, May). America's teachers: Profile of a

profession. NCES 93-025. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Clune, W. H. (1989). The implementation and effects of high school graduation requirements: First steps toward curricular reform. RR-011. New Brunswick, N.J.: Center for Policy Research in Education, Rutgers, The State University of New Jersey.

Coley, R. J. and Goertz, M. E. (1990). Educational standards in the 50 states: 1990. Princeton: Educational Testing Service.

Fetters, W. B., Stowe, P. S., and Owings, J. A. (1984). High School and Beyond: Quality of Responses of High School Students to Questionnaire Items. Washington, D. C.: National Center for Education Statistics.

General Accounting Office (1993, January). Student testing: Current extent and expenditures, with cost estimates for a national examination. GAO/PEMD-93-8. Washington, D. C.: United States General Accounting Office.

Goodwin, D. (1991, August). Beyond defaults: Indicators for assessing proprietary school quality. Washington, D.C.: U.S. Department of Education.

Hoachlander, E. G., Kaufman, P., Levesque, K., and Houser, J. (1992). Vocational education in the United States: 1969-1990. NCES 92-669. Washington, D.C.: National Center for Education Statistics.

Horn, L., Hafner, A., and Owings, J. (1992, June). A profile of American eighth-grade mathematics and science instruction. (National Education Longitudinal Study of 1988). NCES 92-486. Washington, D. C.: U.S. Department of Education, National Center for Education Statistics.

Huelskamp, R. M. (1993, May). Perspectives on education in America. Phi Delta Kappan, 74(9), 718-721.

Jencks, C. (1980). Declining test scores: An assessment of six alternative explanations, Sociological Spectrum, premier issue, 1-15.

Kaufman, P. and Rasinski, K. A. (1991). National Education Longitudinal Study of 1988: Quality of the Responses of Eighth-Grade Students in NELS:88. Washington, DC: US Department of Education, National Center for Education Statistics.

Koretz, D. M. (1986). Trends in Educational Achievement. Washington, D. C.: Congressional Budget Office.

Koretz, D. M. (1987). Educational Achievement: Explanations and Implications of Recent Trends. Washington, D.C.: Congressional Budget Office.

Koretz, D. M. (1990). Trends in the Postsecondary Enrollment of Minorities. Santa Monica: RAND (R-3948-FF).

Koretz, D. M. (1991). State comparisons using NAEP: Large costs, disappointing benefits. Educational Researcher, 20 (3), April, 19-21.

Koretz, D. (1992a). Evaluating and validating indicators of mathematics and science education. N-2900-NSF. Santa Monica: RAND.

Koretz, D. (1992B). What Happened to Test Scores, and Why? Educational Measurement: Issues and Practice, Vol. 11, No. 4, pp. 7-11.

Koretz, D. M., Linn, R. L., Dunbar, S. B., and Shepard, L. A. (1991). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. In R. L. Linn (Chair), Effects of High-Stakes Testing on Instruction and Achievement, symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April 5.

Koretz, D. M., Madows, G. F., Haertel, E., and Beaton, A. G. (1992). Statement before the Subcommittee on elementary secondary and vocational education, Committee on education and labor, U.S. House of Representatives.

Langbein, L. I. and Lichtman, A. J. (1978). Ecological inference. Beverly Hills: Sage.

Lapointe, A. E., Askew, J. M., and Mead, N. A. (1992, February). Learning science. International Assessment of Educational Progress Report No. 22-CAEP-02. Princeton: Educational Testing Service.

Lazar, S. (1992, June). Learning about the world. International Assessment of Educational Progress Report No. 22-CAEP-05. Princeton: Educational Testing Service.

Linn, R. L, and Dunbar, S. B. (1990). The Nation's report card goes home: Good news and bad about trends in achievement. Phi Delta Kappan, 72 (2), October, 127-133.

Linn, R. L, Graue, M. E., and Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of the claims that "everyone is above Average. Educational Measurement: Issues and Practice, 9 (3), 5-14.

Madaus, G. F. and Kellaghan, T. (1991, April). Examination systems in the European Community: Implications for a national examination system in the United States. Contractor Report PB 92-127570. Washington, D.C.: U.S. Congress, Office of Technology Assessment.

McDonnell, L. M. (1993, July). Student assessment as an instrument of education policy. DRU-438-UCLA/OERI. Santa Monica: RAND.

McDonnell, L. M., Burstein, L., Ormseth, T. H., Catteral, J. S., and Moody, D. (1990). Discovering what schools really teach: Designing improved coursework indicators. JR-02. Santa Monica: RAND.

Moody, A. C. and Christoff, D. (1992). A study of U.S. teacher supply and demand: Fifth in a series. Evanston: Association for School, College and University Staffing, Inc.

Mullis, I. V. S., Dossey, J. A., Foertsch, M. A., Jones, L. R., and Gentile, C. A. (1991). Trends in Academic Progress. Washington, D. C.: National Center for Education Statistics.

Mullis, I. V. S., Dossey, J. A., Campbell, J. R., Gentile, C. A., O'Sullivan, C., and Latham, A. S. (1994). *NAEP 1992 Trends in Academic Progress*. Washington, D. C.: National Center for Education Statistics.

Murnane, R. J., and Raizen, S. A. (Eds.) (1988). Improving indicators of the quality of science and mathematics education in grades K-12. Washington, D.C.: National Academy Press.

National Assessment of Educational Progress (1981). Three National Assessments of Reading: Changes in Performance, 1970-1980. Denver: NAEP/Education Commission of the States.

National Assessment of Educational Progress (1985). The Reading Report Card: Progress Toward Excellence in Our Schools. Princeton: NAEP/Educational Testing Service.

National Assessment of Educational Progress (1988a). The Mathematics Report Card: Are We Measuring Up? Princeton: NAEP/Educational Testing Service.

National Assessment of Educational Progress (1988b). The Science Report Card: Elements of Risk and Recovery. Princeton: NAEP/Educational Testing Service.

National Assessment of Educational Progress (1990). The Reading Report Card, 1971-88. Princeton: NAEP/Educational Testing Service.

National Center for Education Statistics (1984). High School and Beyond: Quality of Responses of High School Students to Questionnaire Items. Washington D.C.: U.S. Department of Education (NCES 84-216).

National Center for Education Statistics (1989a). Dropout Rates in the United States: 1988. Washington, D.C.: U.S. Department of Education (NCES-89-609).

National Center for Education Statistics (1989b). Report of the NAEP Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons. Washington, D.C.: U.S. Department of Education (CS-89-499).

National Center for Education Statistics (1994). Digest of Education Statistics. Washington D.C.: U.S. Department of Education (NCES 94-115)

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: author.

National Educational Goals Panel. (1991). Measuring progress toward the national education goals: Potential indicators and measurement strategies. Washington, D.C.: Author.

National Educational Goals Panel. (1993). The national education goals report: Building a nation of learners. Washington, D.C.: Author.

National Educational Goals Panel. (1992). The national education goals report: Building the best. Washington, D.C.: Author.

National Research Council. (1994). National science education standards: Draft for Response and Comment. Washington, D.C.: National Academy Press.

National Science Board. (1993). Science and engineering indicators. NSB 93-1. Washington, D.C.: U.S. Government Printing Office.

Nicholas, G. S. (1992). Teacher supply and demand in the United States: 1992 report. Evanston: Association for School, College and University Staffing, Inc.

Oakes, J. (1986, October). Educational indicators: A guide for policymakers. OPE-01. Santa Monica: Center for Policy Research in Education, RAND.

Pelavin, S. H. and Kane, M. (1990). Changing the odds: Factors increasing access to college. New York: The College Entrance Examination Board.

Shavelson, R. J. (1987, April). Historical and political considerations in developing a national indicator system. Paper presented at the annual meeting of the American Educational Research Association.

Shavelson, R., McDonnell, L., Oakes, J., Carey, N. (1987, August). Indicator systems for monitoring mathematics and science education. R-3570/NSF. Santa Monica: RAND.

Special Study Panel on Education Indicators. (1991, September). Education counts: An indicator system to monitor the nation's educational health. Washington, D. C.: National Center for Education Statistics

Stecher, B. M. (1992). Describing secondary curriculum in mathematics and science: Current status and future indicators. N-3406-NSF. Santa Monica: RAND.

Stecher, B. M. (1995, April). The cost of performance assessment in science. Symposium presented at the annual meeting of the National Council on Measurement in Education.

Valiga, M. J. (1986). The accuracy of self-reports high school course and grade information. ACT Research Report Series 87-1. Iowa City: American College Testing Program.

Weiss, I. R. (1987, November). Report of the 1985-86 national survey of science and mathematics education. (RTI/2938/00-FR) Research Triangle Park, NC: Research Triangle Institute.

Westbury, I. (1992, June-July). Comparing American and Japanese achievement: Is the United States really a low achiever? Educational Researcher. Vol. 21, No. 5, pp. 18-24.