Public Disclosure Authorized

Public Disclosure Authorized

DIRECTIONS IN DEVELOPMENT

16170
Nov. 1996

Monitoring the Learning Outcomes of Education Systems

VINCENT GREANEY
THOMAS KELLAGHAN

DIRECTIONS IN DEVELOPMENT

Monitoring the Learning Outcomes of Education Systems

Vincent Greaney
Thomas Kellaghan

The World Bank
Washington, D.C.

DIRECTIONS IN DEVELOPMENT

Monitoring the Learning Outcomes of Education Systems

Vincent Greaney
Thomas Kellaghan

The World Bank
Washington, D.C.

Contents

Preface vii

1 Nature and Uses of Educational Indicators 1
Educational Indicators 3
Choice of Outcome Indicators 4
Uses of Information from Outcome Assessments 5
Informing Policy 6
Monitoring Standards 7
Introducing Realistic Standards 7
Identifying Correlates of Achievement 8
Directing Teachers' Efforts and Raising Students' Achievements 8
Promoting Accountability 9
Increasing Public Awareness 9
Informing Political Debate 10
Role of National Assessments 10

2 National and International Assessments 12
National Assessments 12
United States 12
England and Wales 15
Chile 17
Colombia 19
Thailand 21
Namibia 22
Mauritius 24
International Assessments 25
International Assessment of Educational Progress 25
International Association for the Evaluation of Educational Achievement 26
Advantages of International Assessments 27
Disadvantages of International Assessments 28

3 National Assessment and Public Examinations 31
Purposes 31
Achievements of Interest 32
Testing, Scoring, and Reporting 33
Populations of Interest 34
Monitoring 34

Contextual Information 35
High-Stakes and Low-Stakes Testing 37
Efficiency 38
Conclusion 38

4 Components of a National Assessment 40
Steering Committee 41
Implementing Agency 42
Internal Agency 42
External Agency 43
Team from Internal and External Agencies 44
Foreign Experts 44
Building Support 45
Target Population 46
Population Defined by Age or Grade 46
Choice of Levels of Schooling 47
Sampling 47
Choice of Population for Sampling Purposes 48
Sample Selection 48
What Is Assessed? 50
Instrument Construction 52
Type of Test 55
Test Sophistication 58
Nonachievement Variables 58
Administration Manual 59
Review Process 59
Administration 60
Analysis 61
Reporting 62
Average Performance of Students in a Curriculum Area 63
Percentage Passing Items 63
Percentage Achieving Mastery of Curriculum Objectives 63
Percentage Achieving Specified Attainment Targets 64
Percentage Functioning at Specified Levels of Proficiency 65
Cost-Effectiveness 65
Conclusion 66

5 Pitfalls of National Assessment: A Case Study 68
Background to the Initiation of a National Assessment in Sentz 68
School System 68
Response to Education Concerns 69
National Assessment of Educational Standards in Sentz 70
Organization 70
Test Development 71
Implementation 72

Analysis of the Case 73
Responses to Assessment 73
Implementation Procedures 74
A Choice to Make 76

References 77

Appendix. National Assessment Checklist 85

Tables

2.1 Proficiency Levels of Students in Grades 4, 8, and 12, as Measured by U.S. NAEP Mathematics Surveys, 1990 and 1992 14
2.2 Percentage of Students at or above Average Proficiency Levels in Grades 4, 8, and 12, as Measured by U.S. NAEP Mathematics Surveys, 1990 and 1992 15
4.1 Specifications for Mathematics Test: Intellectual Behaviors Tested 54
4.2 Distribution of Costs of Components of National Assessment in the United States 66
5.1 Educational Developments in Sentz, 1970–90 69
5.2 Schedule of Activities for a National Assessment in Sentz 72

Boxes

2.1 Atypical Student Samples 28
4.1 Examples of Multiple-Choice Items in Mathematics, for Middle Primary Grades 55
4.2 Example of Open-Ended Item in Mathematics, for Lower Secondary Grades 56
4.3 Dangers of Cultural Bias in Testing 60

Preface

The collection and publication of statistics relating to numbers of schools, numbers of teachers, student enrollments, and repetition rates have for some time been a feature of most education systems. Up to relatively recently, however, few systems, with the exception of those with public examinations, have systematically collected information on what education systems actually achieve in terms of students' learning. This is so even though, as the World Declaration on Education for All (UNESCO 1990b) reminds us, "whether or not expanded educational opportunities will translate into meaningful development—for an individual or for society—depends ultimately on whether people learn as a result of those opportunities."

In response to this consideration, education systems in more than fifty countries, most of them in the industrial world, have in recent years shown an interest in obtaining information on *what their students have learned* as a result of their educational experiences. This interest was manifested either by developing national procedures to assess students' achievements or by participating in international studies of student achievement. It seems likely that the number of countries involved in these activities will increase in the future.

This book is intended to provide introductory information to individuals with an interest in assessing the learning outcomes of education systems. It considers the role of indicators in this process, in particular their nature, choice, and use (chapter 1). A number of approaches to assessing learning outcomes in selected industrial countries (the United States and the United Kingdom) and in representative developing countries (Chile, Colombia, Mauritius, Namibia, and Thailand) are described. Systems of comparative international assessment are also reviewed, and the arguments for and against the participation of developing countries in such assessments are examined (chapter 2).

Some countries already have available and publish information on student learning in the form of public examination results. The question arises: can such information be regarded as equivalent to the information obtained in national assessment systems that are designed specifically to provide data on learning outcomes for an education system? The answer (reached in chapter 3) is that it cannot.

In chapter 4, the various stages of a national assessment, from the establishment of a national steering committee to actions designed to

disseminate results and maximize the impact of the assessment, are described. Finally, in chapter 5, a case study containing numerous examples of poor practice in the conduct of national assessments is presented. The more obvious examples of poor practice are identified, and corrective measures are suggested.

The authors wish to express their appreciation for assistance in the preparation of this paper to Leone Burton, Vinayagum Chinapah, Erika Himmel, John Izard, Ramesh Manrakhan, Michael Martin, Paud Murphy, Eileen Nkwanga, O. C. Nwana, Carlos Rojas, Malcolm Rosier, Molapi Sebatane, and Jim Socknat. The manuscript was prepared by Teresa Bell and Julie-Anne Graitge. Nancy Berg edited the final manuscript for publication. Abigail Tardiff and Amy Brooks were the proofreaders.

1
Nature and Uses of Educational Indicators

Although most of us probably think of formal education or schooling primarily in terms of the benefits that it confers on individuals, government investment in education has often been based on assumptions about the value of education to the nation rather than to the individual. As public schooling developed in the eighteenth and nineteenth centuries, support for it was frequently conceived in the context of objectives that were public rather than private, collective rather than individual (Buber 1963). More recently, colonial administrations recognized the value of education in developing the economy as well as in promoting shared common values designed to make populations more amenable to control.

The importance of education for the nation is reflected in the considerable sums of money that national governments, and, frequently, provincial, regional, and state governments, are prepared to invest in it. In 1987 world public expenditure on education amounted to 5.6 percent of gross national product (GNP); the figure varied from a low of 3.1 percent for East Asia to a high of 6.5 percent for Oceania. As a percentage of total government expenditure, the median share for education was 12.8 percent in industrial countries, a figure considerably lower than the 15.4 percent recorded in developing countries (UNESCO 1990a).

Given this situation, it is not surprising that for some time government departments have routinely collected and published statistics that indicate how their education systems are working and developing. Statistics are usually provided on school numbers and facilities, student enrollments, and efficiency indices such as student-teacher ratios and rates of repetition, dropout, and cohort completion. But despite an obvious interest in what education achieves, and despite the substantial investments of effort and finance in its provision, few systems in either industrial or developing countries have, until recently, systematically collected and made available information on the outcomes of education. Thus, in most countries there is a conspicuous dearth of evidence on the quality of students' learning. Few have stopped, as a former mayor of New York was inclined to do, and asked "Hey, how am I doing?" although knowing precisely how one is doing would surely be useful.

Since the 1980s, however, decisionmakers have begun to attach increasing importance to the development of a coherent system for monitoring and evaluating educational achievement, specifically pupil learning outcomes. In this book, our focus is on the development of such a system. Following usage in the United States, this system is referred to as a national assessment.

The interest in developing a systematic approach to assessing outcomes—in doing a national assessment—can be attributed to several factors. One is a growing concern that many children spend a considerable amount of time in school but acquire few useful skills. As Windham (1992) has pointed out, school attendance without learning "makes no social, economic or pedagogical sense" (p. 56). In the words of the World Declaration on Education for All (UNESCO 1990b, par. 4),

> Whether or not expanded educational opportunities will translate into meaningful development—for an individual or for society—depends ultimately on whether people actually learn as a result of those opportunities, in other words, whether they incorporate useful knowledge, reasoning ability, skills, and values.

The problem of inadequate school learning is not confined to developing countries. Throughout the world, one hears expressions of dissatisfaction with the levels of achievement of today's students, though there may be little evidence that standards are in fact declining. Even without such evidence, a case can still be made that changes in the world of work are resulting in a mismatch between educational outcomes and the needs of society (Townshend 1996). This mismatch is most obvious in the case of what has been called "an educational underclass" made up of students who perform very poorly in the education system. This underclass is found in most countries. In the past its members could find employment in unskilled work, but this is no longer possible because jobs that require only minimal literacy skills are fast disappearing from the labor market, particularly in industrial countries.

Given the need for better-educated students, decisionmakers are concluding that a monitoring system is necessary to gather information needed to describe and monitor the nature of students' achievements, the relevance of those achievements to the world of work, and the number of inadequately prepared students leaving the system.

What is learned at school assumes even more importance because of increased global economic competition, marked by rapid movement of capital and new technologies from country to country. In such a situation, it is claimed that a country's level of productivity and ability to compete depend greatly on workers' and management's skill in using

capital and technology (World Bank 1991) and thus that "skilled people become the only sustainable competitive advantage" (Thurow 1992, p. 520). Comparative studies of students' achievements have been used to gauge the relative status of countries in developing individual skills.

Another reason for interest in monitoring student achievements is that governments today are faced with the problem of expanding enrollments while at the same time improving the quality of education—without increasing expenditure. More detailed knowledge of the functioning of the education system will, it is hoped, help decisionmakers cope with this situation by increasing the system's efficiency.

A final reason for the increased interest in monitoring and evaluating educational provision arises from the move in many countries, in the interest of both democracy and efficiency, to decentralize authority in the education system, providing greater autonomy to local authorities and schools. When traditional central controls are loosened in this way, a coherent system of monitoring is necessary.

Educational Indicators

The term *educational indicator* (in the tradition of economic and social indicators) is often used to describe policy-relevant statistics that contain information about the status, quality, or performance of an education system. Several indicators are required to provide the necessary information. In choosing indicators, care is taken to provide a profile of current conditions that metaphorically can be regarded as reflecting the "health" of the system (Bottani and Walberg 1994; Burnstein, Oakes, and Guiton 1992). Indicators have the following characteristics (Burnstein, Oakes, and Guiton 1992; Johnstone 1981; Owen, Hodgkinson, and Tuijnman 1995):

- An indicator is quantifiable; that is, it represents some aspect of the education system in numerical form.
- A particular value of an indicator applies to only one point or period in time.
- A statistic qualifies as an indicator only when there is a standard or criterion against which it can be judged. The standard may involve a norm-referenced (synchronic) comparison between different jurisdictions; a self-referenced (diachronic) comparison with indicator values obtained at different points in time for the same education system; or a criterion-referenced comparison with an ideal or planned objective.
- An indicator provides information about aspects of the education system that policymakers, practitioners, or the public regard as

important. Sometimes it may be easy to obtain consensus among interested parties on what is important; other times it may not.

- An indicator is realistic in the sense that it is based on information collected with due regard to financial and other constraints.
- An indicator describes conditions amenable to improvement.
- Information for indicators is collected frequently enough to allow change to be monitored.
- Indicators allow an examination of distributions among subpopulations of interest (for example, by age, gender, income, or socioeconomic group).
- The selection of indicators to represent the status of the education system is based on a model, which may be explicit or implicit, of how the education system works (Burnstein, Oakes, and Guiton 1992). The set of indicators incorporated in the model should reflect the multifaceted nature of education in all its complexity (Bottani and Tuijnman 1994) and be comprehensive enough to describe the important dimensions of the system. The model, in turn, provides a context for interpreting what the indicators mean, how they relate to other aspects of the education system (and perhaps to other social and economic systems), and how they are likely to respond to various kinds of manipulation.

The model of the education system on which indicators are built frequently comprises some combination of inputs, processes, and outputs. *Inputs* are the resources available to the system—for example, buildings, books, the number and quality of teachers, and such educationally relevant background characteristics of students as the socioeconomic conditions of their families, communities, and regions. *Processes* are the ways schools use their resources as expressed in curricular and instructional activities. *Outputs* are all that the school tries to achieve; they include the cognitive achievements of students and affective characteristics such as the positive and negative feelings and attitudes students develop relating to their activities, interests, and values.

Choice of Outcome Indicators

To enumerate the outcomes of education about which it might be useful to have empirical information in terms of the many aims that have been posited for education would be an endless task. Aims frequently suggested include the development of literacy and numeracy skills, the development of aesthetic areas of experience, preparation for life in a democratic society, preparation for the world of work, development of character and moral sensitivity, and personal self-fulfillment. Aims (and

expected outcomes) may differ for different ages and students. Given the range of educational aims and the complexity and difficulty of measuring outcomes, some selection has to be made in deciding what outcomes should be measured for use in an indicator system. All we can hope for is information on a limited number of indicators rather than a description of all possible outcomes.

In choosing indicators, the evaluator will be influenced by consideration of which educational outcomes are regarded as important and by the ease and efficiency with which the outcomes can be measured. Thus, both political and technical considerations have to be attended to. At the political level, some measures will be regarded as more important or credible than others, and achieving consensus on these may not be easy. At the technical level, considerations relating to such factors as method of measurement, sampling strategies, and how data are aggregated and reported may also constrain the selection of indicators (Burnstein, Oakes, and Guiton 1992).

The role of both political and technical factors is evident in the emphasis placed on cognitive factors in assessing school outcomes. Partly because it is difficult to obtain agreement on the value as school outcomes of activities with a large noncognitive component, and partly because these activities are difficult to measure, most attention in the development of outcome measures has been given to cognitive achievement. The general public, as well as those professionally involved in education, seems genuinely interested in finding out what cognitive skills and knowledge students have acquired in school. For example, can students read and write satisfactorily? Is their knowledge of science and technology adequate for life in the contemporary world?

In a national assessment, measures of achievement in key curriculum areas are administered to students at selected age or grade levels. The measures used are similar to those frequently used in classrooms to assess students' achievements. However, the purpose of the exercise is not to obtain information on how individual students are performing but to measure, through national aggregation of individual student performances, the achievement of the education system (or of some part of it).

Uses of Information from Outcome Assessments

Information about achievement outcomes provides objective measures of the state, quality, or performance of an education system. This information can be used for a variety of purposes. In this section we consider eight such uses: informing policy, monitoring standards, introducing realistic standards, identifying correlates of achievement, directing

teachers' efforts and raising students' achievements, promoting accountability, increasing public awareness, and informing political debate.

Informing Policy

Information on the achievements of students in an education system can serve a variety of audiences and functions. Educational administrators, such as senior ministry of education officials, should be in a position to produce valid, timely, and useful information when addressing policy issues to be resolved in a political setting. Without such information, policymaking can be unduly influenced by personal biases of ministers of education or senior civil servants, vested interests of school owners or teacher unions, and anecdotal evidence offered by business interests, journalists, and politicians. Given this range of influences, at a minimum, pertinent data must be available to guide the selection of priorities in curriculum, the provision of material resources, and teacher training strategies. However, as noted above, factual information to assist policymaking, especially data on the quality of student learning, is seldom available in developing countries. Even when data on student achievement are available, the views of powerful constituencies will continue to play a role in setting educational priorities. Virtually all decisions in public policy are based on both facts and values (Lincoln and Guba 1981). The role of achievement data is to strengthen the factual basis of decisionmaking.

Many education systems are committed to the principle of equality of opportunity and monitor the extent to which groups enjoy equal access to and participate in education. Information from a national assessment can bring this a step further by providing evidence about the achievements of such groups. Thus, national assessment results have been used in the United States to provide evidence of differences in school achievement related to geography, gender, and ethnicity. Many countries will also be interested in knowing whether mean reading achievement levels are similar for boys and girls, rural and urban children, and children from different linguistic groups.

Information from a national assessment will be more useful to policymakers if it provides information on subdomains of knowledge rather than just an overall score for a curriculum area such as reading or mathematics. Recent reading surveys have examined respondents' performance in analysis and comprehension of narrative material (based on fictional text), expository material (information or opinion writing), and documentary material (information presented in a structured form in charts, maps, lists, or sets of instructions) (Elley 1992). In mathemat-

ics, categories (subdomains) that have been used include numbers and operations, measurement, geometry, data analysis and statistics, and algebra and functions (Lapointe, Mead, and Askew 1992). Data on the performance of students in subdomains can point to strengths and weaknesses within curriculum areas, show how intended curricula are implemented in schools, and, in particular, highlight such factors as gender, urban-rural location, or performance at different times. Such information may have implications for curriculum design, teacher training, and the allocation of resources.

Monitoring Standards

Information on student achievement in key curriculum areas collected on a regular basis has helped monitor changes in achievement over time in such countries as Chile, France, Ireland, Thailand, the United Kingdom, and the United States. By presenting objective findings on achievement, a national assessment can provide evidence relevant to assertions made frequently by employers, industrialists, and others that educational standards are falling.

Countries vary in the frequency with which they obtain information on particular areas of achievement. A five-year interval would seem to be a reasonable time span, since achievement standards are unlikely to vary greatly from year to year. This does not mean that a national assessment exercise would be conducted only every five years. Assessments could be more frequent, but a particular curriculum area would be assessed only once in five years.

Introducing Realistic Standards

A national assessment can foster a sense of realism in the debate on appropriate achievement levels. In developing countries, unrealistic standards have probably contributed to the high student failure rates that are a feature of many education systems (Kellaghan and Greaney 1992). Unduly high levels of expectation may be prompted by the desire to maintain traditional colonial standards. However, such a target may be almost impossible to attain, given the level of socioeconomic development of some countries. Another factor affecting the target is the changing nature of the school-going population arising from the dramatic increase in enrollment numbers; this increase, in turn, is often accompanied by lower teacher qualification requirements and a decline in the quality of educational facilities.

Identifying Correlates of Achievement

Information on correlates of the outcomes of an education system can help policymakers identify factors over which they can exercise some control—factors likely to contribute to improvements in student achievement levels. Data on some of these potentially manipulable variables may have to be collected along with achievement data at the time of the national assessment. For example, national assessment data have been used in Colombia to assess the impact of in-service teacher training. In Chile the contribution of school resources to student achievement has been examined and decisions made about the allocation of such resources. Other possible correlates of achievement include the emphasis placed on individual subject areas; assessment and supervision procedures; textbooks (prices, numbers, contents, and distribution systems); curricular content; and state policies on language instruction.

Directing Teachers' Efforts and Raising Students' Achievements

The expectation is that action will be taken in the light of national assessment results to mandate changes in policy or in the allocation of resources. However, the information such assessments provide may be sufficient, even without formal action, to bring teaching and learning into line with what is assessed (Burnstein, Oakes, and Guiton 1992). The reason for the improvement is that the indicators may point to what is important, and "what is measured is likely to become what matters" (Burnstein, Oakes, and Guiton 1992, p. 410). As a consequence, curricula, teaching, and learning will be directed toward the achievements represented in the indicators. What is tested is what will be taught, and what is not tested will not be taught (Kellaghan and Greaney 1992).

The conditions under which assessments will have positive effects are not entirely clear. Certainly, there are situations in which assessment systems have little impact on policy or practice (Gipps and Goldstein 1983), for example, when the results are not communicated clearly or in a usable way to policymakers. It is equally certain that when high stakes are attached to performance on an assessment, teaching and learning will be aligned with the assessment (Kellaghan and Grisay 1995; Madaus and Kellaghan 1992). But although this may result in improved test scores, if these are the result of teaching to the test, they will not necessarily be matched by improvement in students' achievement measured in other ways (Kellaghan and Greaney 1992; Le Mahieu 1984; Linn 1983).

Thailand provides an example of a national assessment designed to change teachers' perceptions of what is important to teach. The assessment included affective outcomes such as attitudes toward work, moral

values, and social participation in the hope that teachers would begin to stress learning outcomes other than those measured in formal examinations. Subsequently, it was established that teachers began to emphasize affective learning outcomes in their teaching and evaluation (Prawalpruk 1996).

Promoting Accountability

Governments need access to relevant information on the operation of the education system to enable them to determine whether the state is getting good value for its investment. That investment is substantial. Recent figures indicate that in most low-income economies, expenditure on education is one of the largest cost items in government spending—much larger than expenditures on health, defense, housing, social security, or welfare (World Bank 1995a). In this situation, relevant feedback is obviously essential and can help avoid a waste of scarce resources that has been described as socially intolerable, economically unacceptable, and politically short-sighted (Bottani 1990, p. 336).

A variety of models of accountability exists. The precise model employed will depend on many factors. First, it will depend on who is regarded as responsible for performance: the teacher, the school, the ministry of education, or the general public. Second, the nature of the information obtained will affect which individuals or institutions are identified as accountable. In the British system of national assessment, information is available about all schools; thus schools can be identified in the accountability process. If individual teachers or schools are not identified in national assessments, it obviously will not be possible to hold them accountable for student performance. Similarly, when samples, rather than whole populations of schools, are tested in a national assessment, adequate information will not be available (except for a small number of sample schools) to identify and hold accountable poorly performing teachers or schools.

Increasing Public Awareness

Ministries of education are often reluctant to place in the public arena information about the operation of the education system that they regard as sensitive. This is not surprising when the ministry is charged by government with attaining politically sensitive (but practically difficult) objectives such as promotion of a national language. Willingness to publicize policy failures is not a conspicuous characteristic of most ministries. In addition, political expediency may dictate that ministries not

report results which highlight the superiority of particular ethnic, linguistic, or regional groups. In such situations, it may be difficult to establish an atmosphere in which national assessments can be conducted and results made freely available to all interested parties.

Although it may sometimes be in the interest of a ministry to control the flow of information, the long-term advantages of an open-information system are likely to outweigh any short-term disadvantage. Several long-term benefits can be identified. When the results of a national assessment are made widely available, they can attract considerable media attention and thus heighten public consciousness on educational matters. The results of a national assessment can also bring an air of reality and a level of integrity to discussions about the education system. The informed debate that is simulated can, in turn, contribute to increased public support for national, regional, and local efforts to improve the education system. Thus, although the knowledge furnished by national assessments may create immediate problems for politicians and government officials, in the longer term it can provide a stimulus, rationale, or justification for reform initiatives.

Informing Political Debate

National and, even more notably, international comparative assessment exercises give rise to considerable debate among politicians, as well as others interested in education. An education system provides a country with the human resources and expertise necessary to make it competitive in international markets, and from this perspective political interest in national achievement is understandable. Politicians need to know whether the education system is giving value for the considerable portion of the national budget they allocate to it each year. Today, in many countries, rhetoric (usually uninformed) tends to dominate the political debate on education. Armed with objective evidence on the operation of the system, politicians are more likely to initiate reforms and to prompt ministries of education to action.

Role of National Assessments

Although there has been a pronounced increase in recent years in support for formal assessment of student achievement (Lockheed 1992), most developing countries still lack valid and timely information on the outcomes of schooling. A national assessment can help fill this gap by providing educational leaders and administrators with relevant data on student achievement levels in important curricular areas

on a regular basis. These data can contribute to policy and public debate, to the diagnosis of problems, to the formulation of reforms, and to improved efficiency.

There is no single formula or design for carrying out a national assessment. A government's purposes and procedures for assessing national levels of achievement will be determined by local circumstances and policy concerns. The diversity of uses and approaches will become more apparent in chapter 2 when we review seven national assessment systems from different regions of the world, as well as international comparative assessments of student achievements. The remainder of the book provides information on how to—and how not to—conduct a national assessment.

It may seem reasonable to argue that spending money on a national assessment is not justified when resources are inadequate for building schools or for providing textbooks to students who need them. In response, it needs to be pointed out that the resources required for the conduct of a national assessment would not go very far in addressing major shortcomings in the areas of school or textbook provision. Furthermore, the information obtained through a national assessment can bring about cost-efficiencies by identifying failing features of existing arrangements or by producing evidence to support more effective alternatives. However, it is up to the proponents of a national assessment to show that the likely benefits to the education system as a whole merit the allocation of the necessary funds. If they cannot show this, the resources earmarked for this activity might indeed be more usefully devoted to activities such as school and textbook provision.

2
National and International Assessments

National assessments tend to be initiated by governments—more specifically, by ministries of education. International assessments often owe their origin to the initiatives of members of the research community. The main difference between the two types of assessment is that national assessments are designed and implemented within individual countries using their own sampling designs and instrumentation, whereas international assessments require participating countries to follow similar procedures and use the same instruments.

In this chapter, national assessment systems in two industrial countries (the United States and England and Wales) and five developing countries (two in Latin America, one in Asia, and two in Africa) are described. Next, two international assessments are outlined, and the advantages and disadvantages for developing countries of participating in such assessments are considered.

National Assessments

National assessments are now a standard feature of education systems in several industrial countries. The assessments are similar in many ways. Virtually all use multiple-choice or short-answer questions, although Norway and the United States include essay-type writing tasks and oral assessments are conducted in Sweden and the United Kingdom (England, Wales, and Northern Ireland). National assessments also differ in several respects from country to country. In Canada and France many grades are assessed, whereas relatively few are assessed in the Netherlands, Norway, Scotland, and Sweden. The purposes of national assessment also vary.

United States

The U.S. National Assessment of Educational Progress (NAEP) is the most widely reported national assessment model in the literature. It is an on-

going survey, mandated by the U.S. Congress and implemented by trained field staff, usually school or district personnel. The survey is designed to measure students' educational achievements at specified ages and grades. It also examines achievements of subpopulations defined by demographic characteristics and by specific background experience. Since 1990 voluntary state-level assessments, in addition to the national assessments, have been authorized by Congress (Johnson 1992).

Although the NAEP has been in existence since 1969, politicians and the general public appear to have become interested in its findings only recently (Smith, O'Day, and Cohen 1990). Heightened political interest as a result of the attention paid by the National Governors' Association to NAEP findings led to the introduction in 1990 of state-by-state comparisons (Phillips 1991). Over the years, details of the administration of the NAEP have changed—for example, the frequency of assessment and the grade level targeted. At present, assessments are conducted every second year on samples of students in grades 4, 8, and 12. Eleven instructional areas have been assessed periodically. Most recent reports have focused on reading and writing (Applebee and others 1990a, 1990b; Langer and others 1990; Mullis and Jenkins 1990); mathematics and science (Dossey and others 1988; Mullis and Jenkins 1988; Mullis and others 1993); history (Hammack and others 1990); geography (Allen and others 1990); and civics (Anderson and others 1990). Data have been reported by state, gender, ethnicity, type of community, and region.

Up to 1984, the percentages of students who passed items were reported. Since that date, proficiency scales have been developed for each subject area. These scales were computed by using statistical techniques (based on item response theory) to create a single scale representing performance (Phillips and others 1993). The scale is a numerical index that ranges from 0 to 500. It has three achievement levels—basic, proficient, and advanced—at each grade level and allows comparison of performance across grades 4, 8, and 12.

In setting the achievement levels, the views of teacher representatives (sixty-eight in mathematics, for example), administrators, and members of the general public were taken into account (Mullis and others 1993). Performance at the lowest, or *basic,* level denotes partial mastery of the knowledge and skills required at each grade level. For example, grade 4 students performing at the basic level are able to perform simple operations with whole numbers and show some understanding of fractions and decimals. Performance at the middle, or *proficient,* level demonstrates competence in the subject matter. In the view of the National Assessment Governing Board, all students should perform at this level. Grade 4 students who are proficient in mathematics can use whole numbers to estimate, compute, and determine whether results are reasonable; have a conceptual understanding of fractions and

decimals; can solve problems; and can use four-function calculators. The highest, or *advanced,* level indicates superior performance. Grade 4 students who receive this rating can solve complex nonroutine problems, draw logical conclusions, and justify answers.

Average mathematics proficiency marks are presented for grades 4, 8, and 12 for 1990 and 1992 in table 2.1. The data in the last column show that in both years more than one-third of students at all grade levels failed to reach the basic level of performance. However, the figures in this and in other columns suggest that standards rose between 1990 and 1992.

Results based on one common scale (table 2.2) show that most students, especially those in grades 4 and 8, performed poorly on tasks involving fractions, decimals, and percentages. Furthermore, very few grade 12 students were able to solve nonroutine problems involving geometric relations, algebra, or functions. Subsequent analyses revealed that performance varied by type of school attended, state, gender, and level of home support.

Comparisons of trends over time show that achievements in science and mathematics have improved, whereas, except at one grade level, there has been no significant improvement in reading or writing since the mid-1980s (Mullis and others 1994).

Information collected in the NAEP to help provide a context for the interpretation of the achievement results revealed that large proportions of high school students avoid taking mathematics and science courses.

Table 2.1. Proficiency Levels of Students in Grades 4, 8, and 12, as Measured by U.S. NAEP Mathematics Surveys, 1990 and 1992

Grade and year	*Average proficiency*	*Percentage of students at or above*			*Percentage of students below basic*
		Advanced	*Proficient*	*Basic*	
Grade 4					
1990	213	1	13	54	46
1992	218	2	18	61	39
Grade 8					
1990	263	2	20	58	42
1992	268	4	25	63	37
Grade 12					
1990	294	2	13	59	41
1992	299	2	16	64	36

Source: Mullis and others 1993.

Table 2.2. Percentage of Students at or above Average Proficiency Levels in Grades 4, 8, and 12, as Measured by U.S. NAEP Mathematics Surveys, 1990 and 1992

Grade and year	*Average proficiency*	*Percentage at or above proficiency level*			
		200	*250*	*300*	*350*
Grade 4					
1990	213	67	12	0	0
1992	218	72	17	0	0
Grade 8					
1990	263	95	65	15	0
1992	268	97	68	20	1
Grade 12					
1990	294	100	88	45	5
1992	299	100	91	50	6

Note: Skills for each proficiency level are as follows:

Level 200. Addition, subtraction, and simple problem solving with numbers

Level 250. Multiplication and division, simple measurement, two-step problem solving

Level 300. Reasoning and problem solving involving fractions, decimals, percentages, and elementary concepts in geometry, algebra, and statistics

Level 350. Reasoning, problem solving involving geometric relationship, algebra, functions.

Source: Mullis and others 1993.

Among eleventh-graders who enroll in science courses, approximately half had never conducted independent experiments. Almost two-thirds of eighth-graders spend more than three hours a day watching television.

England and Wales

In England and Wales, national monitoring efforts have been a feature of the education system since 1948. Large-scale national surveys of levels of reading achievement of 9-, 11-, and 15-year-olds were conducted irregularly up to 1977 (Kellaghan and Madaus 1982). In 1978, partly in response to criticisms about standards in schools, a more elaborate system of assessment, run by the Assessment of Performance Unit in the Department of Education and Science, was set up (Foxman, Hutchinson, and Bloomfield 1991). Three main areas of student achievement were

targeted for assessment at ages 11, 13, and 15: language, mathematics, and science. In addition to pencil-and-paper tests, performance tasks were administered to small samples of students to assess their ability to estimate and to weigh and measure objects.

Assessments in the 1980s carried considerable political weight. They contributed to the significant curriculum reform movement embodied in the 1988 Education Act, which, for the first time, defined a national curriculum in England and Wales (Bennett and Desforges 1991). The new curriculum was divided into four "key" stages, two at the primary level and two at the secondary level. A new system of national assessment was introduced in conjunction with the new curriculum. Attainment was to be assessed by teachers in their own classrooms by administering externally designed performance assessments. These assessments went well beyond the performance tests introduced by the Assessment and Performance Unit; they were designed to match normal classroom tasks and to have no negative backwash effects on the curriculum (Gipps and Murphy 1994).

The policy-related dimension of the assessments was clear. They were intended to have a variety of functions: *formative*—to be used in planning further instruction; *diagnostic*—to identify learning difficulties; *summative*—to record the overall achievement of a student in a systematic way; and *evaluative*—to provide information for assessing and reporting on aspects of the work of the school, the local education authority, or other discrete parts of the education service (Great Britain, Department of Education and Science, 1988). In particular, the assessments were expected to play an important role in ensuring that schools and teachers adhered to the curriculum as laid down by the central authority . Thus the assessment approach could be described as "fundamentally a management device" (Bennett and Desforges 1991, p. 72); it was not supported by any theory of learning (Nuttall 1990).

Although there have been several versions of the curriculum and of the assessment system since its inception, some significant features of the system have been maintained. First, all students are assessed at the end of each key stage at ages 7, 11, 14, and 16. Second, students' performance is assessed against statements of attainment prescribed for each stage (for example, the student is able to assign organisms to their major groups using keys and observable features, or the student can read silently and with sustained concentration). Third, assessments are based on both teacher judgments and external tests.

Teachers play an important role in assessment: they determine whether a student has achieved the level of response specified in the statement of attainment, record the achievement levels reached, indicate level of progress in relation to attainment targets, provide evidence to support levels of attainment reached, and give information about stu-

dent achievements and progress to parents, other teachers, and schools. Moderation is carried out by other teachers, to help ensure a common marking standard.

Initial reactions to the process indicated that teachers welcomed the materials provided and the innovative assessment procedures. On the negative side, the assessment process placed a heavy burden on teachers, the in-service support provided was inadequate, and the assessment turned out to be largely impractical (Broadfoot and others n.d.; Gipps and others 1991; Madaus and Kellaghan 1993). To add to the problems, results were being published at a time of intense competition between schools and of job losses, which gave rise to questions about entrusting the administration and scoring to teachers (Fitz-Gibbon 1995).

Two important lessons can be drawn from the British national assessment system. First, the use of complex assessment tasks leads to problems of standardization of procedures for administration and scoring that, in turn, lead to problems of comparability, both between schools and over time. Second, it is extremely difficult, if at all possible, to devise assessment tasks that will serve equally well formative, diagnostic, and summative evaluative purposes (Kellaghan 1996c). Efforts to deal with these problems are to be found in the move to make greater use of more conventional centralized written tests and to accord priority to the summative function in future assessments (Dearing 1993; Gipps and Murphy 1994).

Chile

In 1978 Chile's Ministry of Education assigned responsibility for a national assessment to an external agency, the Pontificia Universidad Católica de Chile. The study was piloted over a two-year period. Data on contextual variables, as well as on achievement, were collected (Himmel 1996). These included student-home variables (student willingness to learn, parental expectations for their children); teacher-classroom variables (teaching methodologies, classroom climate); principal and school variables (expectations of staff and of students, promotion of parents in school activities); and institutional variables (educational and financial policies).

The assessment was designed to provide information on the extent to which students were achieving learning targets considered minimal by the Ministry of Education; to provide feedback to parents, teachers, and authorities at municipal, regional, and central levels; and to provide data to planners that would guide the allocation of resources in textbook development, curriculum development, and in-service teacher training.

All students in grades 4 and 8 were assessed in Spanish (reading and

writing), mathematics, and the natural and social sciences. The testing of all students was justified on the grounds that teachers in the project were more likely to react to the results if they considered that their students had contributed to them directly (Himmel 1996). Very small schools and schools in inaccessible locations were excluded. In all, 400,000 students in grades 4 and 8, or approximately 90 percent of the relevant populations, participated.

In 1984 a new minister for education announced that the assessment system was to be abolished. Although the reasons for the change in policy were not announced, it appears that senior officials considered the cost (estimated at $5 per student) too high. An effort to revive the assessment system in 1986 failed because of a lack of technical competence and resources within the ministry. Educational supervisors, however, continued to support the concept. In 1988, with strong support from a new minister, a national assessment was reintroduced under the title Sistema de Información sobre la Calidad de la Educación (SIMCE). Responsibility for the conduct of the assessment was again assigned to the Universidad Católica, with the proviso that after four years project execution would be transferred to the ministry. Objectives were similar to those in the earlier assessment.

Separate teams were established to take responsibility for technical issues (instrument development, analysis, and provision of documentation); information (including preparation of data bases, optical scanning, data processing, and report printing); and administration (including planning, hiring, transportation, and contracting).

Prior to the administration of instruments, an intensive dissemination campaign was launched to help develop a positive attitude toward the SIMCE. The campaign included technical brochures (developed separately for parents and schools), posters for schools, videos for workshops, and a nationwide television and press release program.

Spanish and mathematics tests (forty-five items each) and a writing test were administered to all grade 4 students in November. In addition, as part of a matrix design, sixteen-item tests in natural science and in history and geography were administered to 10 percent of students. Two affective measures—a multiple-choice self-concept test and a questionnaire of student perceptions—were given to all students. Teacher questionnaires were administered to five teachers in each school, and a parent questionnaire to parents of all students.

All instruments were administered during a two-day period and returned to Santiago within fifteen days. Multiple-choice tests were machine scored. Writing tests were hand-scored over a sixteen-day period. All multiple-choice items and 10 percent of open-ended items were scored. Results revealed that students performed poorly in relation to curriculum objectives. Students in urban schools performed better than

students in rural schools; students in large schools performed better than students in small schools; and students in private schools scored highest.

The results were disseminated extensively. Teachers received classroom results containing the average percentage of correct answers for each objective assessed, as well as the average number of correct answers over the entire test. Results were also reported nationally and by school, location, and region. Each classroom and school was given a percentile ranking based on other schools in the same socioeconomic category, as well as a national ranking. Special manuals explained the results and indicated how schools and teachers could use the information to improve achievement levels. Results were given to school supervisors.

Relatively little use was made of the self-concept information. Parental information was not used and was not collected after the first year. Parents, however, received a simplified report of overall results for Spanish and mathematics.

Use of the national assessment results has increased gradually. Low-scoring schools have access to a special fund to enable them to improve infrastructure, educational resources, and pedagogical approaches. Results have also been used to prompt curriculum reform. Percentile rank scores were dropped in favor of percentage scores because teachers found it difficult to interpret the former.

The Chilean experience highlights the need for consensus and political will, technical competence, and economic feasibility (Himmel 1996). Currently there appears to be political and public support for the SIMCE. It provides education administrators with information for planning, and authors of instructional materials use the information to identify objectives. However, the enterprise has not been a total success. Some schools, realizing that their rank depended on the reported socioeconomic grouping of their students, overestimated the extent of poverty among their students to help boost their position. Efforts to explain procedures and results to parents have not been reflected in increased parent involvement with schools except for private schools. Almost two-thirds of teachers reported that they did not use the special manual that dealt with the pedagogical implications of the test results. Finally, questions have been raised about the value of the census approach when sample data could provide policymakers with the needed information.

Colombia

National assessment in Colombia was prompted by a perception that insufficient relevant information was available for decisionmaking at central, regional, and local levels (Rojas 1996). The Ministry of Education also wished to use the results to generate debate on educational issues.

The initial assessment conducted in 1991 focused on the extent to which standards defined as minimum in mathematics and language were being attained in grades 3 and 5 in urban and rural public and private schools. A total of 15,000 students participated in the assessment. Originally thirteen states, accounting for 60 percent of the population, were targeted. The sample comprised 650 students in grade 3 and 500 students in grade 5 in each state.

For grade 3 four performance levels were assessed in mathematics and three in reading comprehension. Performance levels or target standards were determined by the test development personnel. For example, in mathematics the lowest performance level included items on simple addition, whereas more complex tasks involving problem solving were equated with higher performance levels. For grade 5 five performance levels were assessed in mathematics and four in reading. Both multiple-choice items and items for which students had to supply short answers were used. Data on personal, school, and environmental characteristics were collected, as well as information on student participation in local organizations or associations.

The national leader of the assessment had considerable experience in research, data collection, and fieldwork. Teams were established to coordinate the fieldwork within individual states. Each team was led by a coordinator who directed the field testing, supported by two or three individuals with formal qualifications in the social sciences. Local coordinators, usually young people, supervised the work of ten to fifteen fieldworkers. The fieldworkers, often university students or recent social science graduates, administered the tests and conducted teacher interviews. The supply of applicants for these positions was ample because of the relatively high unemployment rates among graduates. Local teachers were not asked to administer tests because it was felt they might attempt to help students taking the tests. Ministry of Education officials were considered unqualified for the work.

At the end of the assessment, profiles of high-scoring schools, teachers, and administrators were developed. The percentages of students who scored at each performance level were reported separately for each state, for public and private schools, and for urban and rural schools, as well as at the national level. Correlates of achievement were identified; these included the number of hours per week devoted to a subject area, teachers' emphasis on specific content areas, teachers' educational level, school facilities, and number of textbooks per student. Negative correlations were recorded for grade repetition, absenteeism, time spent getting to school, and family size (Instituto SER de Investigación/ Fedesarrollo 1994). The number of in-service courses a teacher had taken did not emerge as a significant predictor of achievement.

Results were released through the mass media, and a program of national and local workshops was organized to discuss the results and their implications. Individual teachers received information on national and regional results in newsletters, brochures, and other user-friendly documents. Administrators, especially at the state level, used results for local comparisons. A national seminar used the national assessment data to identify appropriate strategies for improving educational quality. Results for individual schools were not reported because it was felt that this would undermine teacher support for the assessment.

The apparent success of the initial assessment has been attributed to the creation of an evaluation unit within the Ministry of Education; to the commitment of the minister and vice-minister for education; to the support of ministry officials; to the use of an external public agency to design the assessment instruments; and to the use of a private agency to take responsibility for sampling, piloting of instruments, administration of tests, and data analysis (C. Rojas, personal communication, 1995). After the first two years responsibility for the national assessment was transferred to a public agency, which administered the assessment in 1993 and 1994. By late 1995, however, the agency had not managed to analyze the data collected in either year.

Thailand

Following the introduction of a new higher secondary school curriculum in 1981, public certification examinations at the end of secondary school were abolished in Thailand, and teachers were given responsibility for evaluating student achievements in their respective courses. Concerned that achievement might fall in this situation, the Ministry of Education introduced national assessment as a means of monitoring standards (Prawalpruk 1996). Administrators at various levels of the system were expected to use the results to help improve the quality of education. To encourage schools to broaden their objectives and instructional practices, the national assessment included measures of affective learning outcomes (attitudes toward work, moral values, and participation) and practical skills.

Starting in 1983, all grade 12 students (in their final year in secondary school) were assessed in Thai, social studies, and physical education. In addition, science, mathematics, and career education were assessed in most subsequent years. Both cognitive and affective outcomes were assessed in social studies, physical education, and career education. The task was entrusted to the Office of Educational Assessment and Testing Services in the Department of Curriculum and Instruction Development.

Many of the staff had achieved master's degrees in educational assessment; eight had been trained outside Thailand. Subject matter committees (twelve to eighteen members each) established for each subject area developed tables of specifications for achievement and wrote multiple-choice items. Nationwide testing was conducted on the same two days.

Schools were furnished with individual student scores and with school, regional, and provincial mean scores; information on how other individual schools performed was not provided. For public communication purposes, student performance was reported as the percentage of items answered correctly. Provincial administrators advised how the results could be used in planning academic programs at school, provincial, and regional levels.

In subsequent years, samples of grades 6 and 9 were assessed, generally every second year. In a reaction to the initial failure of schools to use assessment results to improve school practice, the national assessment design was expanded to include measures of school process (school administration, curriculum implementation, lesson preparation, and instruction). Starting in 1990 school process measures were assessed by teams of three external evaluators. The early national assessment results for science and mathematics were considered disappointing; they showed that students were weak at applying principles in both subject areas. This conclusion prompted a significant curriculum revision in 1989.

National assessment has been used for school and provincial planning and for monitoring levels of student achievement over time; it has also helped increase teacher interest in affective learning outcomes. According to Prawalpruk (1996), some principals misused the results by claiming that poor results could be attributed to poor teaching. Results were used for educational planning only if adequate administrative support was available. School principals ignored assessment results if they did not consider them useful for planning.

Namibia

The National Institute for Educational Development in Namibia collaborated with Florida State University and Harvard University in 1992 to assess the basic language and mathematics proficiencies of students at grades 4 and 7. The objectives of the assessment were to inform policymakers on achievement levels to enable them "to decide on resource targeting to underachieving schools" (Namibia, Ministry of Education and Culture, 1994, p. xiv), to sensitize managers to the professional needs of teachers, to enable schools and regions to

compare themselves with their counterparts, and to provide baseline data for monitoring progress.

Tests were developed "by reference groups within the head office of the ministry" (p. 7), based on official curricula and textbooks. A random sample of 136 schools was drawn, covering Namibia's six education regions. Within each school, one grade 4 and one grade 7 class were chosen randomly. In one specific region of interest (Ondangwa), thirty-four schools with grade 4 students and nineteen with grade 7 students took the national language—Oshindonga—test. Test instructions to all students were given in the local language. More than 7,000 students in grades 4 and 7 were tested in English and mathematics.

Of the 136 schools, 20 were included in a special longitudinal sample to monitor changes in English achievement over time. In these schools, students in grades 4 and 5 took the grade 4 test, whereas those in grades 6 and 7 took the grade 7 test. It was planned to readminister the tests to students each year. It is now accepted that the longitudinal sample was too small to permit generalization to the wider population of Namibian children.

The tests were administered to all students in attendance in the targeted grades in the 136 sample schools; only 98 schools, however, had a grade 7 class. Both the English and Oshindonga tests were timed. The English test took 40 to 60 minutes and the Oshindonga test 60 to 80 minutes to complete. The untimed mathematics test took up to 120 minutes and caused some student fatigue.

Because the test designers hoped to get a normal distribution of test scores, tests were not designed to assess levels of mastery. Items answered correctly by less than 20 percent or more than 80 percent of students were deleted in analyses. This reduced severely the number of items that could be used in measuring performance levels—in the English grade 4 test, from seventeen to nine, and in the grade 7 mathematics test, from sixty to thirty-eight.

Results showed that many grade 4 students had difficulty with the English test, prompting concern that the expected level of performance was too high and suggesting that the curriculum materials might be too advanced. Initial analyses of results suggested that most categories of students increased their scores between grades 4 and 5 and between grades 6 and 7. At grade 7, the performances of girls and boys were similar on the two language tests, but boys outscored girls on the mathematics test. Older students had much lower scores than younger ones; for example, 19-year-olds answered correctly fewer than half the items answered correctly by 12- and 13-year-olds on both the English and mathematics tests. Differences in scores for regions and for language groups were also reported.

Data were used to relate performance levels to three background factors—age, gender, and home language—which in combination explained about one-third of the variance in English scores and about one-fifth of the variance in mathematics scores. In one region, however, less than 3 percent of the variance could be attributed to these factors. A set of papers was prepared for teachers outlining practical suggestions for improving student performance in areas that had posed difficulties.

The study concluded that the process of developing the tests for the assessment was not altogether satisfactory and that a new competency-based curriculum will make it necessary to develop new measures to assess basic competencies in subject areas.

Mauritius

To implement the recommendations of the World Conference on Education for All, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) and the United Nations Children's Fund (UNICEF) launched a project to develop national assessment capacities in China, Jordan, Mali, Mauritius, and Morocco (Chinapah 1992; UNESCO 1994). Identification missions to each country were supported by some centralized training in survey methodology. Each national assessment focused on learning achievement (literacy, numeracy, and basic life skills); factors related to learning achievement (personal characteristics, home environment, and school environment); and access and equity (female enrollment, and admission and participation rates of specific groups). The designers hoped that lessons learned in the course of the project could be adapted and applied in other developing countries.

The national assessment in Mauritius was conducted to address policy issues relating to educational inequalities (Chinapah 1992) and to provide baseline data on achievement levels, with the aim of identifying the percentage of students who attained defined acceptable standards in specified subject areas. Literacy (English and French), numeracy, and life skills were assessed. Items on road safety, awareness of the environment, social skills, and study skills were included.

Specific performance criteria were developed for each subject area (Mauritius Examinations Syndicate 1995). To be rated literate in French, for example, a 9-year-old was required to obtain a minimum score of twenty marks out of thirty-five, including eight of a possible thirteen in "reading" and twelve of twenty-two in "vocabulary, written expression." To be considered literate in English, the 9-year-old was expected to obtain a minimum score of seventeen marks, including twelve out of a possible twenty-two in reading and five of eight in writing. Such performances were considered to represent the ability to read clearly, to un-

derstand different types of text judged appropriate for 9-year-olds, and to solve simple shopping problems (V. Chinapah, personal communication, 1995).

Approximately 1,600 standard IV students, mainly 9-year-olds in a representative sample of fifty-two schools, were assessed. Questionnaires were administered to parents, teachers, and school principals to obtain background information on home, school, and student characteristics. Responsibility for the assessment was entrusted to the Mauritius Examinations Syndicate. The syndicate, which administers the annual high-stakes public examinations, had some technical competence in test development, data analysis, and administration of formal assessments. Each test lasted 40 minutes. The literary and numeracy test relied on multiple-choice and short-answer questions, the life skills test on multiple-choice items. Tests were administered by retired primary school inspectors and head teachers. Data were collected in 1994, and findings were presented to the Ministry of Education and to teachers. The syndicate plans to repeat the assessment in the future to monitor possible changes in achievement over time (R. Manrakhan, personal communication, 1995).

International Assessments

International assessments, in contrast with national assessments, involve measurement of the educational outcomes of education systems in several countries, usually simultaneously. Representatives from many countries (usually from research organizations) agree on an instrument to assess achievement in a curriculum area, the instrument is administered to a representative sample of students at a particular age or grade in each country, and comparative analyses of the data are carried out (Kellaghan and Grisay 1995).

Countries participating in international studies are expected to provide personnel and funds for administration, training, printing, local analyses, and production of national reports. Costs of instrument development, sampling frameworks, international data analyses, and report writing are the responsibility of the international assessment agency, to which individual countries make a financial contribution.

International Assessment of Educational Progress

The first International Assessment of Educational Progress (IAEP), conducted in 1988 under the direction of Educational Testing Services, under contract to the U.S. Department of Education, represents an

extension of the U.S. 1986 NAEP assessments in mathematics and science. The IAEP involved 13-year-olds in five countries and four Canadian provinces (Lapointe, Mead, and Phillips 1989). Items selected from the original pool of items used in the NAEP were adapted to take account of cultural differences. The second IAEP project, conducted in 1991, was much more extensive. The mathematics and science achievements of 9- and 13-year-old students were assessed in twenty countries (Lapointe, Askew, and Mead 1992; Lapointe, Mead, and Askew 1992). Data on a broad array of contextual variables, including time given to homework, availability of books in the home, teacher characteristics, extent of urbanization, and teaching approaches, were also obtained.

International Association for the Evaluation of Educational Achievement

The International Association for the Evaluation of Educational Achievement (IEA), headquartered in The Hague, has been carrying out studies of school achievement, attitudes, and curricula in a variety of countries since 1959. Studies require an elaborate test development process in which participating countries are invited to contribute, review, and pretest items. Although one of the IEA's primary functions is to conduct research designed to improve understanding of the educational process (Visalberghi 1990), the association was also intended to have a more practical and applied purpose: to obtain information relevant to policymaking and educational planning in the interest of improving education systems (Husén 1987; Plomp 1993).

To date, the IEA has conducted studies of mathematics achievement (Husén 1967; Travers and Westbury 1989) and of science achievement (Comber and Keeves 1973; Keeves and Rosier 1992; Postlethwaite and Wiley 1992). In language, it has carried out studies of reading literacy (Elley 1992), written composition (Gorman, Purves, and Degenhart 1988), English as a foreign language (Lewis and Massad 1975), and French as a foreign language (Carroll 1975). It has also conducted investigations of civic education (Torney, Oppenheim, and Farnen 1976), computers in education (Pelgrum and Plomp 1991), and preprimary childcare (Olmstead and Weikart 1989).

The studies have amassed a substantial body of information on a range of educationally relevant variables. Levels and patterns of achievement in a variety of curricular areas have been described and compared across countries. So also have differences in intended and implemented curricula and in the course-taking patterns of students. A variety of correlates of achievement have been identified, including students' opportu-

nity to learn, the amount of time a subject is studied, the use of computers, and factors and resources in the homes of students (Anderson and Postlethwaite 1989; Anderson, Ryan, and Shapiro 1989; Elley 1992, 1994; Kifer 1989; Lambin 1995; Postlethwaite and Ross 1992).

Advantages of International Assessments

The main advantage of international studies over national assessments is the comparative framework they provide in assessing student achievement and curricular provision (Husén 1967). International assessments give some indication of where the students in a country stand relative to students in other countries. They also show the extent to which the treatment of common curriculum areas differs across countries, and, in particular, the extent to which the approach in a given country may be idiosyncratic. This information may lead a country to reassess its curriculum policy.

Many accounts are available of how findings of international studies on student achievement and curricula have been used to change educational policy (Husén 1987; Kellaghan 1996b; Torney-Purta 1990). For example, results of international studies have been credited with the increased emphasis placed on science in Canada and in the United States (McEwen 1992). In Japan the relatively superior performance of students in mathematical computation compared with mathematical application and analysis led to a change in emphasis in the curriculum (Husén 1987). In Hungary participation in IEA studies has been credited with curriculum reform in reading, and the finding that home factors accounted for more variance in student achievement than school factors helped to undermine Marxist-Leninist curricular ideologies (Báthory 1989).

International assessments have many other advantages. Their findings tend to attract more political and media attention than those of national studies. Thus, poor results can provide politicians and other policymakers with a strong rationale for budgetary support for the education sector.

For national teams entrusted with the implementation of international assessment, the experience of rigorous sampling, item review, printing, distribution, supervision, scoring, data entry, and drafting of national reports according to an agreed-on timetable can contribute greatly to the development of local capacity to conduct research and national assessments. Finally, staffing requirements and costs are lower in international studies than in national assessments because instrumentation and sampling design are developed in collaboration with experts in other countries.

Disadvantages of International Assessments

It can be argued that such factors as availability of schools and materials, opportunity to learn, status and quality of teaching, parental interest, and class size differ so radically from country to country that valid comparisons of international achievement test results are impossible (Rotberg 1991). Although IEA studies generally consider the extent to which students in individual countries have had opportunities to learn the content tested, it is doubtful whether politicians, policymakers, or the media take these into consideration when commenting on national rankings. Political rhetoric, frequently based on the perceived implications of the findings for competitiveness in international trade rather than on a sober evaluation of the meaning of results, may dominate the discussion immediately following the publication of results. In fairness, it should be stressed that uninformed political rhetoric can be prompted by the results of national as well as international assessments and that some of the problems associated with international assessments apply equally to national assessments.

A potentially significant problem with both international and national studies is the difficulty in obtaining a representative sample of students (box 2.1). In many developing countries up-to-date population data may not be available, and communication and logistical problems can contribute to relatively low response rates. The National Center for Education Statistics in the United States has set a response rate target of 85 percent for cross-sectional surveys. This target may be much too high for developing countries, and indeed it has been achieved only once by the United States in international studies of mathematics and science (Medrich and Griffith 1992). Sampling problems are commonplace and have been blamed for significant reversals of performance in some countries between grades (Rotberg 1991). Targeted populations may not be comparable, especially in countries where national enrollment, drop-

Box 2.1. Atypical Student Samples

In the 1991 IAEP mathematics study, only 3 percent of the population of 13-year-old students in Brazil and 1 percent of the corresponding population of students in Mozambique were sampled. The performance of Chinese students—which was highlighted in the report of the study—was based on a sample that excluded many 13-year-olds: those below grade 7 in twenty provinces and cities, those out of school (almost 50 percent of the population), and those attending school in nine provinces and autonomous regions with predominantly non-Chinese populations (Lapointe, Mead, and Askew 1992). The exclusion of these groups suggests that the reported achievement levels may seriously overestimate the mean achievements of Chinese students.

out, and retention rates differ sharply. The result is that countries may have been represented by atypical samples of students.

A further problem with international assessments is that it is probably impossible to develop a test that is equally valid for several countries (Kellaghan and Grisay 1995). What is meant by "achievement in mathematics" or "achievement in science" varies from country to country because different countries will choose different skills applied to different facts and concepts to define what they regard as mathematical or scientific achievement. Furthermore, a particular domain of a subject may be taught at different grade levels in different countries. For example, simple geometric shapes, which are introduced in many countries in the junior or lower primary grades, are not introduced until grade 5 in Bangladesh. Again, prior knowledge or expectations might interfere with attempts to solve a simple problem.

Because items included in an international test represent a common denominator of the curricula of participating countries, it is unlikely that the relative weights assigned to specific curriculum areas in national curricula will match those in international tests. In the 1988 IAEP relatively little effort was made to test the curricula covered by non-U.S. participants. As a result, in one of the participating countries (Ireland), important areas of the mathematics curriculum were not tested, and other areas that received substantial emphasis in the national curriculum were accorded relatively little emphasis in the international test (Greaney and Close 1989).

Although a range of test formats is used in international assessments, the multiple-choice format is used widely for reasons of management efficiency and desirable psychometric properties (especially reliability). Even when other assessment formats are included, reports may be limited to the results of the multiple-choice tests. This means that important skills in the national curriculum, including writing, oral, aural, and practical skills, are excluded.

The costs of international assessments are likely to be lower than those of national assessments, but participation in an international assessment does require considerable financial support. The IEA estimates that the minimum national requirement is a full-time researcher and a data manager. Personnel requirements vary according to the nature of the assessment. Developing countries that wish to participate must pay a nominal annual fee and make a contribution to the overall costs on the basis of their economic circumstances. Local funds have to be obtained for printing, data processing, and attendance at IEA meetings. Costs may be met by a ministry of education, from university operating budgets, or from a direct grant from the ministry of education to a university or research center. IEA experience suggests that government-owned institutes have a better track record than universities in conducting assess-

ments (W. Loxley, personal communication, 1993). A lack of meaningful contact between university researchers and government ministries is particularly noteworthy in some Latin American countries.

Many developing countries are likely to encounter a range of common problems, whether they are conducting an international or a national assessment. These include unavailability of current population information on schools and enrollment figures; lack of experience in administering large-scale assessments or in administering objective tests in schools; tests that do not adequately reflect the curriculum offered in schools or that fail to reflect regional, ethnic, or linguistic variations; lack of exposure to objective-type items; fear that test results might be used for teacher accountability purposes; insufficient funds and skilled manpower to do rigorous in-country analyses of the national or international data; governmental restrictions on publicizing results; and logistical problems in conducting the assessment.

On balance, a developing country can probably benefit from participation in international assessments of student achievements. Participation can help develop expertise that can be drawn on later in more focused and more relevant national assessments. Consultant support, however, may be needed to carry out an international or national assessment. In particular, the services of long- and short-term local and foreign consultants may be required to offer training programs in test development, sampling, and analysis.

3
National Assessment and Public Examinations

Although the idea of national assessment is new in most countries, public examinations are an important and well-established feature of education in Africa, Asia, Europe, and the Caribbean. In developing countries they are usually offered at the end of primary schooling and at the ends of the junior and senior cycles of secondary schooling. Public examinations are similar in many respects to national assessments: procedures are formalized, and testing is normally done outside the classroom setting and requires students to provide evidence of achievement. Because of their importance, their frequency, and their similarity to national assessments, it is reasonable to ask whether public examinations could be used to obtain the kind of information that national assessment systems are designed to collect.

Eight issues are relevant in attempting to answer this question: the purposes of public examinations and of national assessments; the achievements of interest to the two activities; testing, scoring, and reporting procedures; the populations of interest to the two activities; monitoring capabilities of the two activities; the need for contextual information in interpreting assessment data; the implications of attaching high stakes to assessments; and efficiency and cost-effectiveness in obtaining information.

Purposes

The purposes of public examinations and national assessments are significantly different. The purpose of a public examination is to determine whether an individual student possesses certain knowledge and skills. A national assessment is not primarily concerned with identifying the performance of individual students; rather, its purpose is to assess the performance of all or part of the education system. Given this difference, we can still ask whether it is possible to aggregate the data from individual assessments in public examinations to obtain information on

Note: For a more extended treatment of this topic, see Kellaghan (1996a).

the system. To answer that question, we have to consider the more specific purposes of individual assessment and the implications of these purposes for the kind of assessment procedure used.

In public examinations, information on student performance is used to make decisions about certification and selection, with selection tending to be the more important function (Kellaghan and Greaney 1992; Lockheed 1991). As a consequence, the assessment procedure or examination will attempt to achieve maximum discrimination for those students for whom the probability of selection is high. This is done by excluding items that are easy or of intermediate difficulty; if most students answered an item correctly, the item would not discriminate among the higher-scoring students. However, tests made up solely of more difficult questions will not cover the whole curriculum or even attempt to do so. The result is that public examinations may provide information on students' achievements on only limited aspects of a curriculum.

The purpose of national assessment is to find out what all students know and do not know. Therefore, the instrument used must provide adequate curriculum coverage. From a policy perspective, the performance of students who do poorly on an assessment might be of greater interest than the performance of those who do well.

Achievements of Interest

There is some overlap in the student achievements identified as important by public examinations and national assessments. During the period of basic education, both certification and national assessment are based on information about basic literacy, numeracy, and reasoning skills. If we look at primary certificate (public) examinations, we find that many focus on a number of core subjects, and a glance at several national assessments indicates that they do the same. For example, students' knowledge of a national language and mathematics is included in all national assessment systems.

However, no national assessment attempts the coverage found in public examinations at the secondary level, when students tend to select and specialize in subject areas. The subjects offered vary from one examination authority to another, but it is not unusual to find syllabi and examinations in twenty, thirty, or even more subjects.

National assessments have focused on cognitive areas of development. Thailand (Prawalpruk 1996) and Chile (Himmel 1996) are among the relatively small number of education systems that have attempted to assess affective outcomes. There is now talk in some countries of extending assessments to students' values, attitudes, and aspirations, which are not assessed directly in public examinations. There is also talk of

assessing higher-order and transferable cognitive skills that might apply across a range of curricular areas. If these developments take place, they will have the effect of further separating the common areas of interest of public examinations and national assessments because public examinations are likely to remain subject-bound.

Testing, Scoring, and Reporting

Testing procedures for national assessment differ from those for public examinations in several important ways. First, the quality and structure of tests differ in the two kinds of assessment. In all testing some standardization in procedure is required, if performances are to have a comparable meaning for different students, in different places, and at different times. Public examinations often appear relatively unstructured. Students may be free to write extended essays, and scoring procedures are often not clearly specified and rely heavily on the judgments of individual markers.

A second important area of divergence between tests used in national assessments and in public examinations lies in the content coverage of the tests. Students may be free to choose the particular subset of questions they elect to answer in a public examination. As already noted, extensive content coverage is not required to produce selection tests that will predict later student performance. Even when content coverage is broad enough to meet the requirements of a test used for certification, the coverage cannot be as thorough as that required in national assessment tests, if for no other reason than that it would place an unduly heavy burden on examination candidates. A national assessment, in contrast, should provide a detailed picture of all important areas of the curriculum, even if all students do not respond to all items, if it is to be useful in indicating particular strengths and weaknesses in that curriculum, as reflected by students' test scores.

Scoring and reporting in public examinations usually follow norm-referenced procedures. A crucial topic of interest in a selection test is how a candidate performs with reference to other candidates. The same kind of norm-referencing is often implicit in how certification results are reported. The main information conveyed by a grade of B (or equivalent mark) is not that the student has acquired a particular body of knowledge or skills but rather that he or she has performed better than students who were awarded C or D grades. For national assessments, however, we want to be able to say something about the level of knowledge and skills of students. Because of this, the tests used tend to be criterion-referenced rather than norm-referenced, and results are often reported in terms of certain performance criteria—for example, that a

certain percentage of students can perform some mathematical operation or has reached a defined level of proficiency in a curriculum area.

Populations of Interest

If public examinations were to be useful for national assessment, they would have to provide information for those populations of students of interest to policymakers and administrators. However, although the first public examination is usually not held before the end of primary schooling, most national assessments obtain information on students at an earlier stage in their educational careers. Thus, there is a consensus that information is required before the age at which students sit for a public examination.

The reason for the consensus is clear. Information for national assessment should lead to decisions designed to improve the quality of education. Because the foundations of later achievement are laid in primary school, it is important to know whether student achievement is poor at this stage so that remedial action can be taken. We also know that grade repetition and dropout are serious problems in many countries during primary schooling (Lockheed, Verspoor, and associates 1991; World Bank 1995b). In these situations, information obtained at the end of primary schooling will come too late for effective action to be taken.

If information is needed while children are still in primary school, policymakers have to either institute public examinations during primary school or set up a national assessment program. Public examinations, however, would not be a cost-effective way to get this information if the examination results were are not also required for selection or certification. More cost-effective procedures in the form of national assessments are available to monitor levels of achievement in the education system.

Monitoring

Monitoring is an important aspect of national assessment. To be able to say, for example, that student achievements are improving over time (perhaps as a result of educational reforms), information must be obtained at different times. Can public examination data be used for this purpose?

Any information obtained from public examinations about standards over time will be limited to those students who take the examinations and to the subjects that they take. Public examinations are voluntary. Some students may decide against taking them, those who do may

select different subjects, and students within a subject area may choose different questions to answer. For these reasons, public examinations are unlikely to provide information about common achievements for a complete population of students.

If we accept this situation, can we at least be confident that monitoring of the performance of those students who take public examinations would provide reliable information on changes in standards over time, even for limited populations? There are two issues: one relates to changes in the characteristics of the population taking examinations over time, and the second to changes in examinations.

As educational provision continues to expand, and as more students sit for public examinations, the characteristics of examinees change over time. As participation rates increase, the average level of achievement of students tends to fall in a variety of school subjects (Keeves 1994; Willmott 1977). However, this might not be reflected in examination results, in which the percentage of students who pass an examination or obtain a particular grade often remains fairly constant over several years.

In addition to changes in the populations taking public examinations, changes occur in the public examinations themselves from year to year. Examination papers are usually released after they have been taken and are used as guides for later examinees. Unless there is a clear definition of the standards to be maintained when a new examination is constructed, meaningful comparisons about performance from one examination to another will not be possible. Tests for national assessment, in contrast, are not usually made public. Parts of national assessments may be publicized so that schools know what is expected in the tests, but other parts are not released and can be used again. Because complete tests are not released, it is easier to build equivalent examinations, which facilitates comparison of performance over time.

Contextual Information

There are several reasons why information other than information on student achievement should be obtained in national assessments. First, the quality of resources, people, and activities in school is important in itself. Second, because only a small range of educational outputs can be measured, the use of contextual information may prevent schools from placing undue emphasis on the outputs measured to the exclusion of other important factors. And third, by allowing an examination of the interactions of inputs, processes, and outputs, contextual information may provide clues to policymakers about why schools obtain the outcomes that they do (Oakes 1989).

Several kinds of contextual information are likely to be of value in providing clues to policymakers about the determinants of achievement, particularly determinants that might be alterable through changes in educational policy (Lockheed 1991; Messick 1984; Oakes 1989; Odden 1990). It is interesting to know what students bring to school from their family and community backgrounds that may contribute to successful or poor school performance. Many studies point to the importance of these background factors for school learning (Kellaghan 1994; Kellaghan and others 1993). One cannot assume, however, that factors found to be important in one cultural context will be of similar importance in other contexts. For example, the fairly consistent finding from industrial countries in Europe and North America that family size is negatively related to educational achievement has not been found in studies in Kenya (Bali and others 1984) or in Tanzania (Drenth, van der Flier, and Omari 1983). In these countries a positive rather than a negative relationship was found between family size and educational achievement, measured in a variety of ways (by standardized tests of ability and public examinations at primary and secondary levels).

A second area of contextual information, and one that will be more relevant to decisions about the distribution of educational resources, is the extent to which schools provide access to various subject areas and to diverse skills. In concrete terms, we can ask about the physical facilities in schools, the range of curricula offered, and the availability of learning-support materials such as libraries and laboratories. There is evidence that variation in provision in these areas is more closely related to educational achievement in developing countries than in industrial countries (Levin and Lockheed 1991). It is also important to know *how* school resources are used. It is one thing to have a library or science laboratory; whether the facility is used extensively by students is another matter. The less-material aspects of schools are also important—in particular, instructional leadership and the institutional pressure that the school exerts to get students to work hard (Cohen 1987). Finally, because teachers must be regarded as the key component in any education system, information should be obtained on the professional teaching conditions in a school that may help or hinder teachers in implementing instructional programs.

In theory, contextual information could be collected in conjunction with public examinations. However, public examinations are not designed or administered in the context of an overall model of the education system. Although this, of course, would not preclude the collection of contextual information, collection of the additional data would place an enormous burden on examination authorities, who are often already greatly overstretched. Furthermore, it would not be cost-effective to col-

lect and process information for all individuals and schools taking public examinations.

High-Stakes and Low-Stakes Testing

An examination or test is said to have high stakes attached to it when sanctions or rewards are linked directly to performance. Attaching high stakes to performance on tests, whether public examinations or national assessments, has important consequences. Students, teachers, and curricula are affected in many ways: curriculum and teaching revolve around the examinations, students and teachers put considerable effort into test preparation, and potential low scorers may be prevented from taking the examination to boost the school's overall performance (Madaus and Greaney 1985). High-stakes tests may also affect the validity of measurement through the test corruption and test score pollution that seem to accompany them (Greaney and Kellaghan 1996). If students are taught in such a way that the match between instructional processes and test items is very close, drawing inferences about students' actual skills and knowledge becomes extremely difficult. The problem is particularly acute if we want to test whether a student can apply skills and knowledge to solve new problems, since in this case the problems must be new to the student and not ones taught in class (Haladyna, Nolan, and Haas 1991; Kellaghan and Greaney 1992; Linn 1983).

For public examinations, the high stakes are obvious because the examinations determine future educational and occupational options. High stakes can also be attached to a test even if sanctions are not explicitly associated with individual student performance. If results are used to rank school districts or schools, the tests will be perceived by schools as an important indicator of what is to be valued in education (Madaus and Kellaghan 1992). What is examined will be taught; what is not examined will not.

High stakes are not usually associated with national assessments. In the United States, for example, national assessment provides an unobtrusive measure of the education system, focusing on describing what students know and can do. It does not try to influence directly what goes on in schools. In contrast, high stakes are closely associated with national assessments in Britain, where—in addition to describing what students know and can do—an express goal is to influence directly what goes on in schools.

Whether high stakes are attached to a national assessment is a matter of choice, but it is a choice that should not be made without serious consideration. On the one hand, if no sanction is attached to performance,

the assessment may have no effect on what goes on in schools. On the other hand, if sanctions are attached, the effect on the measuring instrument may be such that improvement in test performance cannot be accepted as evidence of real improvement in the knowledge and skills of students.

Efficiency

National assessments differ from public examinations in three important ways in their implications for efficiency. First, whereas public examinations are held annually, the frequency of national assessments varies from once every year (in Colombia, France, and the United Kingdom) to once every ten years (in Finland). Once every four or five years in a subject area would seem a reasonable compromise and should provide adequate monitoring information because overall achievement standards tend to change slowly.

Second, not every student has to take a test in a national assessment. All that is required is a sample of students that adequately represents the total student population and is large enough for proposed analyses to yield valid and reliable information for policymakers.

Third, it is not necessary for every student who participates in a national assessment to respond to all items. The use of matrix sampling, in which a total test is divided into several components, means that comprehensive content area coverage can be achieved without placing an undue burden on individual students.

Only the last of these issues, it should be noted, would preclude the use of public examinations for national assessment on the grounds of efficiency. If other conditions were satisfactory, national assessment data could be extracted from public examination data for a sample of students at appropriate intervals (for example, every third or fifth year) in a cost-effective way.

Conclusion

One might think of a number of ways to modify public examinations to provide information for a national assessment. A public examination used for certification might be expanded to provide adequate curriculum coverage—although this might have adverse effects on the examination system by, for example, making examinations too long. The emphasis on norm-referencing in public examinations would remain a problem, but it too could possibly be dealt with. As far as the population of interest is concerned, information on students who are too young to

take public examinations could be obtained by introducing a public examination in the primary years. This would not be cost-effective, however, and the introduction of public examinations at an early stage in the educational process might not be beneficial to students' education. Contextual information to assist in the interpretation of student performance could be collected in conjunction with a public examination, perhaps on a sample basis.

Other issues, however, appear to create insuperable problems in the use of an examination system for national assessment. These include the use of public examinations primarily for selection; the difficulty of using them for monitoring standards; and their use to drive instruction in a high-stakes context. Public examinations lack the basis for comparability required for monitoring: examination populations change over time in an unknown way, and methods of scoring are often not known or cannot be demonstrated to be sufficiently consistent. Problems of comparability can emerge even under the highly standardized conditions under which national assessments are conducted.

Finally, there are several reasons why national assessment should not be associated with the high stakes that are attached to public examinations. Such an association would be likely to lead to negative effects on teaching and learning, and, as we saw, if the instrument of measurement becomes corrupted, this is likely to defeat the very *raison d'être* of national assessment, which is to obtain an accurate picture of student achievement in the system.

4
Components of a National Assessment

The sequence of activities associated with national assessment is described in this chapter. Although the list is not meant to be exhaustive or mandatory, it includes activities that national assessment planners should consider.

A national assessment exercise is unlikely to have significant impact if stakeholders in the education system do not agree that its results are likely to have an effect on policymaking concerns. The attitude of the government is particularly important. If a national assessment is to be introduced and subsequently institutionalized within the education system, it should address issues of concern to the ministry for education.

Information on student achievement in an education system can be obtained either through participation in a comparative international study or through an independent national assessment. The nature, relevance, and usefulness of the information obtained will vary depending on the course a country pursues. In an international study, the procedures followed will be determined at the international level. This is necessary because comparisons between countries cannot be made unless the studies follow similar procedures. A national study permits greater flexibility but makes international comparisons impossible; the benefit of learning from other countries' experience is also lost.

When a country decides to do a national assessment, planners should not design one that is overambitious in subjects covered, assessment procedures, sample complexity, or demands on personnel. Almost inevitably there will be tension between the ideal of carrying out a comprehensive assessment and the need to use the initial exercise to develop local capacity in conducting such an exercise. Keeping the scope of an assessment manageable—by, for example, limiting it to one subject and one grade level—increases the chances of a successful operation. Another way to keep an assessment manageable is to limit its geographic coverage. Particularly in large, diverse countries such as China, India, and Indonesia, valuable experience and useful policy-related information can be obtained from assessments confined to one or a few regions of a

Note: An earlier, shorter version of this chapter is presented in Greaney (1996).

country. However, if data from national assessments are to be used to monitor achievement trends over time, limitations in the data-gathering procedures in the early stages will affect the ability to make comparisons in later years, if the procedures change.

Steering Committee

A good way of addressing the political dimension of a national assessment is for the ministry of education to establish a politically heavyweight steering committee. Such a committee would have several functions. It would provide status for the national assessment; it would help ensure that the needs of the powerful national groups in the educational establishment, especially those of the ministry for education, are addressed; and it would help remove the administrative and financial stumbling blocks that can jeopardize or paralyze an assessment effort. The steering committee would also promote public awareness and the discussion of results, thereby maximizing the impact of the national assessment on educational policymaking.

Because the educational–political power structures of countries differ, the interests represented in a national steering committee will vary from country to country. In the United States, for example, the NAEP steering committee included two governors, three professors of education, a school superintendent, and a teacher. In some developing countries representatives of important ethnic, religious, and linguistic groups might be included. Representation should be provided for those responsible for administering the national assessment, those who will consider the results for policymaking, those responsible for funding the exercise, and those who will be entrusted with the policy reforms that may arise from the assessment, such as school administrators and teachers. Addressing the information needs of these stakeholders should help ensure that the exercise does not result in a report that is criticized or ignored because of its failure to address the "correct" questions.

The inclusion of key groups in the steering committee will help neutralize opposition to the exercise. In a more positive vein, those included can provide insights that the committee might otherwise overlook. A strong steering committee can help key stakeholders appreciate the constraints under which other agencies, including the ministry of education, operate. The committee can also help open doors for the implementing agency and ensure that cooperation is forthcoming from the stakeholders. Finally, a steering committee will have a sense of ownership over any proposed reforms that might emanate from the exercise, thereby increasing the likelihood that subsequent policy initiatives will be accepted.

The steering committee should address several important issues at an early stage. These include identifying the purpose and rationale of the national assessment, deciding on the content and on the grade levels to be targeted, developing a budget and assigning budgetary control, selecting an agency or agencies to conduct the assessment, determining terms of reference, and deciding on reporting procedures and publication. The steering committee is likely to be most active at the start of the assessment exercise. The implementing agency will be responsible for most of the detailed work in instrument development, sampling, administration, and reporting, but the steering committee should be provided with draft copies of instruments and descriptions of proposed procedures so that committee members can see that the information needs which prompted the assessment in the first place are being adequately addressed.

Implementing Agency

Two main options are available when it comes to assigning responsibility for conducting the day-to-day work of a national assessment: a ministry of education can do the work itself or it can contract out the work to an external body. A compromise situation, in which a ministry collaborates with an external agency, is also possible. In some cases the assistance of foreign experts may be required. Whatever method is chosen, the implementing organization must have a reputation for competence: it should be able to provide evidence of quality work, technical skills, and integrity, and it will need expertise or access to expertise in project management, research design, curriculum analysis, test and questionnaire development, sampling, printing and distribution, data collection, processing and analysis, and reporting. A perception of competence is important in gaining admission to schools and in getting key individuals and organizations to respond to questionnaires and requests for interviews.

Internal Agency

Many ministers for education may wish to look no further than their own personnel to conduct a national assessment. In Thailand, for example, the Office of Educational Assessment and Training, a section of the Ministry of Education, was given responsibility for the national assessment. In many developing countries some of the most knowledgeable educators may be employed within the ministry. Ministry person-

nel are also likely to have ready access to up-to-date information for sampling purposes, and school inspectors or members of curriculum or textbook units should have considerable insight into key aspects of the education system. In support of the use of an internal agency, ministers might argue that no extra budgetary allocation would be required because the national assessment could be charged to the ministry's operating budget.

There are, however, strong arguments against a ministry of education carrying out a national assessment on its own. Many ministries lack the required technical competence in such areas as instrument development, sampling, and data analysis. A ministry may be put under pressure to withdraw its employees from the assessment to tackle "more pressing" issues in response to the government's educational and political priorities, thus subjecting the assessment to frequent delays. Trying to subsume the cost of a national assessment under current ministry expenditure can also lead to underestimation of costs. For example, the opportunity costs of ministry personnel delegated to work on the national assessment, or other costs such as printing, travel, and data processing, may not be included in the assessment budget.

Another likely difficulty when a ministry carries out a national assessment is that many ministries of education, in both developing and industrial countries, may be slow to share information with others in education or with the public. For example, results that point to poor delivery of an education service or to failure by the formal education system to achieve a particularly sensitive goal (such as equality of achievement for ethnic groups) can embarrass ministry officials and, even more critically, their political masters. Ministry of education staff, compared with external agency personnel, have a greater vested interest in the outcomes of the assessment. As a consequence, they might be less enthusiastic about focusing on potentially awkward issues or in making unpalatable findings public, thereby limiting the effectiveness of the assessment for policy change.

External Agency

A strong case can be made for assigning the responsibility for conducting a national assessment to an external agency. The main stakeholders in education may consider the information provided by a respected nongovernmental or independent agency more objective and thus more acceptable. Also, technical competence, especially in instrument development, sampling, psychometrics, and data processing, is more likely to

be found within university departments and independent research institutes than within ministries of education. For such reasons the national assessment in Chile was assigned initially to a nongovernmental body. In Colombia the Ministry of Education asked its test development center to develop the achievement measures and contracted with a research institute to conduct the other aspects of the assessment.

The use of an external agency, however, is not without its problems. In anticipation of these, a memorandum of agreement should be drawn up between the steering committee and the implementing agency before work begins. The memorandum should deal with such issues as funding, timetables, relations between the two bodies, and permitted data use. When university personnel are entrusted with the assessment, care has to be taken to avoid an overacademic treatment of the data and issues; relevant key policy issues must remain paramount.

It may be advisable to write penalty clauses into contracts with external agencies to help ensure that assessments are completed on time. However, adequate funding must be provided at the outset to complete the task; unlike government departments, research agencies may have little flexibility once the budget has been used up.

Team from Internal and External Agencies

An alternative to sole reliance on an internal or an external agency is to entrust the national assessment to a team composed of both ministry of education personnel and outside technical and curriculum experts. Such an arrangement can capitalize on the strengths of both groups and increase the likelihood of general acceptance of the assessment findings. For example, developmental work for the Mauritian national assessment was undertaken by the semiautonomous Mauritian Examination Syndicate in collaboration with other national agencies, including the Ministry of Education, Science, and Technology and the Institute of Education and its Curriculum Development Center. Both the Examination Board and the Institute of Education are external bodies. UNESCO also provided support.

The team approach can give rise to administrative problems, however, and it may lack stability. Administrative problems may be anticipated by assigning clearly defined tasks and responsibilities to individuals.

Foreign Experts

When the necessary professional competence is not available locally, foreign experts or technical assistance may have to be employed. Data

analysis for the Namibian national assessment was directed by Harvard University, while Florida State University provided assistance with aspects of sampling. Foreign experts should answer to the steering committee. In such instances the development of local capacity to conduct future national assessments should be a priority.

The temptation to entrust a national assessment to a foreign expert or agency should be resisted because this is the scenario most likely to lead to inattention to local capacity-building. Also, policymakers and others involved in the educational enterprise may ignore the findings on the assumption that foreigners or nonresidents are unfamiliar with the local educational context. There is, for example, little evidence to indicate that the assessment of reading achievement in Latvia (Dedze 1995), conducted from Sweden, has had any impact on educational policy within the country.

Building Support

Attitudes toward participation in a national assessment will vary across countries. Where there is a strong tradition of high-stakes assessments in the form of public examinations, teachers may fear—with some cause—that a national assessment will be used as an instrument of accountability. Some national assessments do indeed carry with them elements of accountability involving teachers and schools. It is more difficult to marshal support for such assessments than for ones that do not hold schools or teachers answerable for student performance. Even when national assessments are free of accountability, the assessment may arouse suspicions because of the novelty of the exercise. These suspicions can be allayed by providing interested parties with detailed information about the procedures and expected outcomes of the study. In the Chilean project, for example, principals, teachers, and parents were informed of the assessment and its purpose. The organizers also distributed sample tests and gave some 100 talks to interested parties to explain the project.

To ensure that a national assessment reaps long-term dividends, a broad consensus about objectives should exist among key stakeholders. Findings relating to policy issues should also be reported early enough to affect the policy dialogue. Throughout the periods of instrument development, data gathering, analysis, and reporting, the power structure within the education system should be recognized, and lines of communication with key interest groups, such as the school inspectorate or its equivalent, school managerial authorities, teachers' representatives, and teacher training authorities, should be kept open.

Target Population

Population Defined by Age or Grade

Inferences about the outcomes of an education system are based on an assessment of the achievements of students in the system. To make such inferences, however, it is not necessary or desirable to administer the same test to students of all ages and grade levels. The first decision to be made in a national assessment is whether the target population will be defined by age or grade level (or both). The second is which levels will be targeted for assessment. How the sample that takes the test is chosen within the age or grade levels selected is discussed in the next section.

In some national assessments (for example, Chile and Scotland), only grade is taken into account in defining the population. Many national and international assessments, however, use both student age and grade in their definition. For example, in the IEA literacy study, two populations were defined: students in the grade level containing the most 9-year-olds and students in the grade level containing the most 14-year-olds (Elley 1992). In recent years in the United States the grade level of the majority of students of a particular age has also been selected (Johnson 1992). This strategy can be justified because in industrial countries, especially those with policies of automatic promotion at the end of each grade, the link between grade and age is pronounced.

In developing countries, especially in Latin America and in francophone West Africa, the link between age and grade may not be close because of widely differing ages of entry into school and policies of nonpromotion. In this situation students of similar age will not be concentrated in the same grade. It is not uncommon to find age ranges of up to six years in the senior grades of primary school, as well as students of the same age spread across six grade levels. Choosing a population on the basis of age in such school systems would be disruptive because it would require students from several grade levels to take the tests at the same time. It would also be difficult to identify appropriate test content for such students.

A strong argument can be made in the light of these considerations for targeting grade level rather than age in national assessments in developing countries. In addition, the concept "grade" tends to be associated with a relatively clearly defined section of the national curriculum and with the content of prescribed textbooks. Thus, assessments can be focused on what teachers teach and what students are expected to learn.

The results of grade-targeted national assessments have the potential to guide curriculum and textbook reform and to provide relevant information for in-service teacher education.

Choice of Levels of Schooling

Policymakers want information on the knowledge and skills of students at selected points in their educational careers. In practically all countries that carry out national assessments, students in the primary grades are targeted. In most of these countries national assessments are also conducted at some point at secondary school level, usually at the lower or junior-cycle secondary grades when education is still compulsory. Information at both levels can be valuable. Assessments at the primary school level can identify deficiencies at an early point in the education system that indicate a need for remedial action. Information gained toward the end of compulsory schooling, or at a point when a large proportion of young people is still attending school, can also be useful if it provides some indication of how well students are prepared for life after school. In many developing countries this will be at the primary school level.

The ministry of education's information needs will dictate the precise age or grade level for which information will be collected. If reading comprehension standards are of interest, the early grades—when students are acquiring prereading and initial reading skills—are probably inappropriate. If an education system has low retention rates, the assessment should be done before adolescence to ensure that a larger number of students and teachers are affected by any policy changes prompted by the assessment.

Sampling

Unless a national assessment has been designed to provide information on individual students, teachers, or schools, as in the British national assessment (Gipps 1993) and Latin American country assessments (Horn, Wolff, and Velez 1991), not all students need to be assessed. In most countries a sample rather than a whole population is selected for the assessment. Several factors favor this approach, including reduced costs in gathering and analyzing data, greater speed in data analysis and reporting, and greater accuracy because of the possibility of providing more intense supervision of fieldwork and data preparation (Ross 1987).

Choice of Population for Sampling Purposes

Because the aim of a national assessment is to obtain an estimate of the achievements of students in the system, it might seem appropriate to define the population as *all students in the country at a particular grade or age level.* In practice, however, this is not possible, for two reasons. It is unlikely in a developing country that a central agency such as the ministry of education would have a list of all students attending school in the country. And even if such a list existed, it would not be efficient or feasible to assess a sample of these students because, if randomly chosen, they would be spread over a large number of schools, making data collection difficult and expensive. Because of these conditions, *schools* (or clusters of students) are usually identified as the population to be sampled first.

In some countries complete lists of schools may not be available, or the lists may be too old to be useful. When lists are available, they should be checked carefully to see that all the schools actually exist. Experience has shown that, especially in remote areas, schools may exist on official registers but not "on the ground." Corrupt politicians and administrators sometimes connive to create nonexistent or phantom teachers and schools, which are then supported by national funds.

Problems also arise when schools exist but are not listed in official statistics, as sometimes happens with private schools—in Pakistan, for example. Ingenuity may be required if such schools are to be identified for possible inclusion in a national assessment. One possibility is to first select school administrative areas and then, with local help, to identify unlisted or private schools in the areas.

Another question to be faced in deciding on the population is that of excluding some schools or students from the assessment. It may, for example, be decided that some students are unassessable because they have learning difficulties or because they have limited proficiency in the language in which the assessment will be conducted. Schools attended by such students should be excluded from the population. Because of the cost of data collection, schools in isolated areas may also be excluded, or a decision may be made to group small adjacent schools as one sampling unit. Exclusions should be kept to a minimum. Information on exclusions should be provided in the report of the national assessment.

Sample Selection

Once the population of all schools eligible for selection has been identified, the next step is to select the schools in which students will be assessed. A variety of strategies is available for doing this, and the organi-

zation carrying out the national assessment may need external technical assistance in choosing the strategy and the number of schools (and students) that are most appropriate. Great care has to be taken in this step because inadequate sampling can make it difficult, if not impossible, to make valid statements about national achievement levels of students.

In sampling schools it is common to stratify the population of schools according to variables of interest, such as location (area of the country, urban or rural); type (public or private); ethnic group membership; and religious affiliation. There are two reasons for this: stratification makes for greater efficiency, and it helps ensure that there are sufficient schools and students in the various categories, such as urban and rural, so that differences between schools (and students) in these categories can be examined. To achieve sufficient numbers of schools and students, oversampling within some strata rather than just selecting a sample size proportionate to the number of schools (or students) in the stratum may be necessary. When strata are oversampled, a system of weighting will have to be used in aggregating student scores to ensure that the contribution of groups to overall statistics is proportionate to their size in the total population, not their size in the sample.

Following the selection of schools, the next important decision relates to how students within a school are to be selected for assessment. A decision will already have been made by the national steering committee about the grade or age level of the students who will take part in the assessment. But the decision still has to be made whether all students in a school at the relevant age or grade level will be assessed, or, if there is more than one class at the relevant level, whether one class will be randomly selected or a group of students selected from all classes. Although the assessment of intact classes has obvious administrative advantages, the selection of students from several classes will provide a better estimate of the achievements of students in the school if the students have been assigned to classes according to different criteria or follow different curricula.

It is not necessary that all students take the same test in a national assessment. A technique known as matrix sampling permits the coverage of an extensive array of items by administering different sets of items to different students. In the United States, for example, samples of students are administered one-seventh of the total number of test items developed for each grade. Matrix sampling was also used in Chile. Such sampling permits the coverage of much larger sections of the curriculum and may prove less time-consuming than administering the same test to all students. However, the technical and logistical requirements of printing many different forms of a test, packaging and administering them, and combining test results may be daunting, especially in a country's first national assessment.

Another decision that has to be made is how to deal with nonresponse, caused either by a school's refusal to cooperate in the assessment or by student absence on the day of testing. The report of the national assessment should include a description of how these and other decisions are reached, as well as the number of schools and students involved.

When a complex sampling design involving such procedures as stratification, clustering, and weighting is used to select participants for a study, the effects of the design on sampling error have to be taken into account in analysis of the data collected. Otherwise, the estimate of true sampling variability will be biased.

What Is Assessed?

Both political and technical considerations affect the identification of the knowledge and skills to be examined in a national assessment. The role of political factors is evident in the need to select content that addresses the informational requirements of key policymakers, usually the ministry of education. Technical considerations are apparent in deciding what is technically possible to measure and in evaluating cost and logistical requirements.

Tensions are likely to emerge between the informational needs, goals, and ambitions of those commissioning a national assessment and the ability of the assessment to accommodate them, given financial, technical, administrative, and time constraints. Enthusiasm about using an assessment to gather information on a broad range of issues, however, has to be tempered by realities such as a lack of appropriate tests, limited funds for printing and administering tests and questionnaires, lack of personnel to undertake data analysis, and lack of computer equipment, software, and expertise. This situation imposes on national assessment planners the obligation to be practical and to confine the assessment to what is regarded as essential.

The planners will be facilitated in their task by developing a model of the education system and how its components interact. The model will help identify the important questions the assessment is designed to answer and the analyses required to provide answers to those questions. The model should also help counteract a tendency sometimes found to collect as much data as possible with a view to possible analysis at a future date, a situation that usually leads to too much data, inadequately exploited. One developing country, for example, undertook an overambitious national assessment and was then unable to manage or analyze the massive amount of material that flowed into its ministry of education. The expensive, futile exercise became known as "the warehouse assessment."

The likelihood of installing a national assessment as a feature of the education system is increased when the scope of the initial assessments is kept simple and manageable. Clarity of purpose at the outset can help avoid unnecessary expense in developing, pretesting, printing, distributing, and scoring lengthy questionnaires—much of whose content will never be analyzed or reported, to the dismay of policymakers and steering committee members.

Information normally collected in national assessments can be divided into three main categories. First, all assessments measure *cognitive outcomes of instruction*—specifically, subject matter competence. Second, inspired by evidence that attitude and interest contribute to student learning, many national and international assessments collect data on *affective outcomes* such as attitudes toward specific subjects, values, and levels of interest. Third, most national assessments also collect contextual information on *background variables* such as school and nonschool factors that may contribute to student achievement.

Cognitive outcomes. All countries that conduct national assessments examine the students' first language and mathematics. Science is sometimes included and, in a smaller number of countries, a second language, art, music, and social studies (Kellaghan and Grisay 1995). The attention to language and mathematics is an indication of the importance of these subjects for basic education and merits serious consideration by any country embarking on a national assessment.

A national assessment in Mauritius collected information on achievements in word knowledge, reading comprehension, and English as a second language (Chinapah 1992). The grade 4 national assessment in Chile examined all students in Spanish, writing, and mathematics, whereas only 10 percent of the students were assessed in natural science, history, and geography (Himmel 1996).

In the United States future assessments in language are likely to differ substantially from current practice, which has tended to emphasize reading, vocabulary, and comprehension skills (University of Illinois 1993). Consistent with recent developments in the study of language, assessments will probably focus on how well students can speak, listen, write, and read. Administrative and manpower considerations, however, suggest that it may be some time before developing countries, and indeed many industrial countries, can devote the necessary resources to detailed national assessments of this kind.

Affective outcomes. The Colombian national assessment included measures of student attitudes to school, subjects, and teachers (Rojas 1996). In Chile questionnaires were administered to assess student self-image and self-esteem; questionnaires were also administered to parents and teachers to get their opinions on aspects of the educational system. However, interpretation of affective outcomes in aggregate form proved

problematic (Himmel 1996). At the international level, the IEA reading literacy study evaluated student attitudes to reading by examining the extent of voluntary reading of books, comics, and newspapers, book reading preferences, and the amount of encouragement students received to read and use the library (Elley 1994).

Background variables. In Thailand an attempt was made to identify factors related to the scholastic achievements of grade 3 students, such as socioeconomic status, school size, grade repetition, and access to preschool (Chinapah 1992). In Chile data on the socioeconomic background of students were collected (Himmel 1996). In Colombia data were collected on student gender, educational history (including grade repetition), time devoted to homework and watching television, home background factors, and teacher characteristics such as level of formal education and the amount of in-service training received (Rojas 1996).

The primary concern of a national assessment is to collect data that provide information on the extent to which important goals of the official curriculum are being achieved. This kind of information is most likely to be used by planners, the providers of pre-service and in-service teacher training, the school inspectorate, members of curriculum and textbook bodies, and teachers.

Instrument Construction

The technical aspects of developing assessment instruments (multiple-choice tests, written assignments, practical exercises, oral and aural tests, and attitudinal scales) are beyond the scope of this book (see Bloom, Madaus, and Hastings 1981; Izard 1996; Linn 1989). Some general points may be made, however. First, the content of instruments must be consistent with the overall objective of the assessment. Second, a persistent focus on policy objectives is required. When the objective of an assessment is to measure competence in mathematics (expressed as computational, conceptual, and problem-solving skills) of final-year primary school students, the assessment design must ensure that each of the three elements of mathematics is assessed, analyzed, and reported.

Once terms of reference have been agreed to, instrument construction can begin. The implementing agency is likely to be charged with this task. When documents that specify a curriculum or syllabus are available, they will likely be used to guide decisions on what the assessment instrument will include. However, such documents usually do not provide sufficient detail or indication of the relative importance of topics to allow the agency to develop an instrument without further input.

Subject matter specialists will normally do much of the basic work of developing instruments. Such specialists should not be limited to uni-

versity personnel. Classroom teachers have much to offer from their insights into current practice, their familiarity with teaching priorities, and their knowledge of the curriculum as implemented in the classroom. An important aspect of the subject matter specialists' work will be to select the individual curricular areas that will be included in the assessment instrument and to decide on the weight that will be given to each area.

Experience indicates that a table of specifications can greatly facilitate the test development process (Bloom, Madaus, and Hastings 1981). A typical table consists of a horizontal axis that lists the content areas to be assessed (for example, aspects of the mathematics curriculum for a given grade level) and a vertical axis that presents in a hierarchical arrangement the intellectual skills or behaviors expected of students. Cells are formed at the intersections of the two axes. It is the responsibility of test developers to assign test items or questions to each cell based on their perceptions of the relative importance of the objective represented by the cell. Cells are left empty if the objective is considered inappropriate for a particular content area.

The table of specifications for a mathematics test (table 4.1) is based on a mathematics curriculum for the middle grades of primary school. Separate subtests were designed to measure students' abilities to carry out basic computations, to understand mathematical concepts, and to solve problems. For example, the cell formed by the intersection of the content area, "fractions," and the intellectual behavior, "ability to solve routine problems," represents the objective, "ability to solve routine problems involving fractions." A committee of subject matter specialists, which included teachers, decided to devote five items to this objective. The cell that contained items testing the ability to carry out operations with whole numbers received the highest weighting (twenty-five items). Many cells had no items. The relative weights of importance assigned to each objective guided test development and later the compilation of the final version of the test.

Countries with established traditions of public examinations will expect that test papers will be available to teachers and students after the assessment has been conducted. In a national assessment, however, because the exercise is in principle repeated at a later date to monitor trends in achievement over time, the same test (or a substantial proportion of it) will have to be used again. In this situation, the assessment instrument should not be made public, and all copies should be collected immediately after test administration. The difficulty in preserving test security under prevailing conditions in many developing countries, however, should not be underestimated (Greaney and Kellaghan 1996). Examples of items used in the assessment procedure may be made public so that school personnel know what is expected of students.

Table 4.1. Specifications for Mathematics Test: Intellectual Behaviors Tested

Intellectual behavior	*Content area*						*Overall total*
	Whole numbers	*Fractions*	*Decimals*	*Measurement*	*Geometry*	*Charts and graphs*	
Computation							
Knowledge of terms and facts	1			2			3
Ability to carry out operations	25	4	8				37
Total	26	4	8	2	0	0	40
Concepts							
Understanding of math concepts	1	4			2		7
Understanding of math principles	4	1			2		7
Understanding of math structure	7	2	5				14
Ability to translate elements from one form to another	2			3			5
Ability to read and interpret graphs and diagrams	4		1	2			7
Total	18	7	6	5	4	0	40
Problem solving							
Ability to solve routine problems	14	5	5	3			27
Ability to analyze and make comparisons	2					4	6
Ability to solve nonroutine problems	2						2
Total	18	5	5	3	0	4	35

Type of Test

Most national and international assessments rely to a considerable extent on the multiple-choice test format. A multiple-choice item usually consists of a statement, direction, or question followed by a series of alternative answers, one of which is correct (for examples, see box 4.1).

The advantages of multiple-choice questions include speed of response, ease of marking or correcting, objective scoring, potential for covering a considerable portion of a content area, and high reliability or consistency (Frith and Macintosh 1984). With the advent of sophisticated optical scanning machines, multiple-choice answer sheets can be scored and processed quickly. It is also easy to provide detailed feedback on characteristics of individual items, objectives, and levels of achievement of students classified by, for example, grade, age, gender, and ethnic, linguistic, regional, or religious affiliation.

On the negative side, construction of good multiple-choice tests is time-consuming. It requires a detailed analysis of curriculum documents, textbooks, and teaching practices to identify instructional objectives and relative weights of importance among different units of the curriculum and different cognitive levels (for example, recall of facts, interpretation, analysis, and synthesis). Multiple-choice tests cannot assess important aspects of the curriculum such as oral fluency, writing, and practical skills. Such tests have also been criticized for overemphasizing

Box 4.1. Examples of Multiple-Choice Items in Mathematics, for Middle Primary Grades

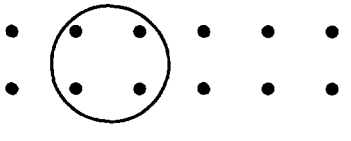
OBJECTIVE: Ability to carry out operations with whole numbers

$$\begin{array}{r} 425 \\ -357 \\ \hline \end{array}$$

(a) 68 (b) 132 (c) 68 (d) Not given

OBJECTIVE: Understanding of fraction concepts

What fraction of the dots is ringed?



(a) 1/2 (b) 1/3 (c) 1/4 (d) 1/5

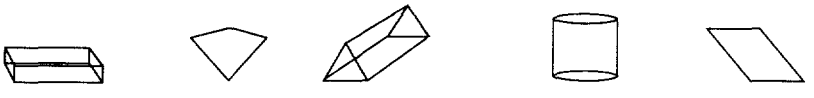
the factual at the expense of determining the student's understanding of the content being assessed.

Criticism of multiple-choice items points to the need to supplement their use with other forms of assessment. In the United States there is a realization that standardized multiple-choice tests have been overused (Cizek 1991; Darling-Hammond and Lieberman 1992). Organizations such as the International Reading Association and the National Council of Teachers of Mathematics have urged the use of alternative assessment approaches. Instead of requiring a student simply to identify a correct answer among a number of possible answers, an item may require the student to construct a response, frequently in the form of a word or a phrase. Usually the answer is a response to a direct question or to an incomplete statement (box 4.2).

Another alternative to the multiple-choice item is a performance task designed to assess competency in such areas as practical measurement skills in mathematics, or the ability to conduct a scientific experiment or to cultivate a plot. National performance testing has value when the emphasis placed on practical skills in national curricula is not reflected in classroom practice or in public assessments such as primary school–leaving certificate examinations. There is evidence that skills ignored in national examinations tend to be neglected in teaching and learning (Kellaghan and Greaney 1992); inclusion of a practical element in a national assessment signals schools that the knowledge and skills involved in practical work are important. Ideally, a practical assessment should provide information on the procedures that students use, their ability to use implements, and the quality of a completed product.

Box 4.2. Example of Open-Ended Item in Mathematics, for Lower-Secondary Grades

OBJECTIVE: Knowledge of geometric shapes



(A) (B) (C) (D) (E)

Which shapes have:

An acute angle? _____
At least one perpendicular line? _____
A rectangular face? _____
At least five corners? _____

One factor inhibiting the wider use of performance tests is that the cost of their administration is considerably higher than the cost of conventional multiple-choice assessments because of the high labor input required. Cost can be reduced, however, by careful selection of a small representative sample of the targeted population. In the United Kingdom, the Assessment of Performance Unit administered practical assignments in mathematics to a sample of approximately 1,000 students as part of its assessments during the 1980s (Foxman and others 1980a, 1980b; Foxman, Ruddock, and McCallum 1990). In these assessments, students were required to weigh and measure objects as part of a series of practical tasks.

Oral and aural tests represent a particular form of performance assessment. In oral assessment, the emphasis is on students' use of spoken language. One index of such use is fluency, ratings of which may be obtained in one-on-one situations. Aspects of oral fluency, such as command of vocabulary and use of idiom, may also be assessed in a structured discussion. Most tasks in the British Assessment of Performance Unit assessment involved both the production and the interpretation of sustained talk (Gipps and Murphy 1994). Whereas assessment of oral competence usually involves individual testing, estimates of aural competence can be obtained in large groups in conventional examination settings; tasks are prerecorded and presented on tape and students write their own responses. Competency in a second language and in music can also be assessed this way.

Interest in assessing students' writing skills has increased greatly in recent years, especially since the introduction of the process approach to writing instruction (Graves 1983). A better test of writing skills is achieved by requiring students to write a series of short passages than by having them write one relatively long passage. As part of a national assessment, students might be asked to write a paragraph about a particular event or a set of instructions or to describe a common object. Another way of assessing students' writing is to have their portfolios (a collection of their writing during the school year) rated by independent assessors using predetermined criteria. Assessments using this method are, however, time-consuming and expensive. Another problem with this kind of assessment is that different scoring methods, such as holistic scoring as against separate scoring of individual elements, yield different results.

Experience with performance assessment has highlighted several problems that arise when it is used in national assessments, including cost, the narrowness of the range of tasks involved, test security, subjectivity in scoring, and the fact that the performance test may not yield comparable data from teacher to teacher or from school to school (Mehrens 1992; Meisels, Dorfman, and Steele 1995). Evidence from Brit-

ain also indicates that performance assessments conducted by teachers as part of a national curriculum assessment yield different results from externally administered tests of closely related content areas (Gipps and Murphy 1994). Given these problems, it would seem reasonable to delay the introduction of performance tasks into national assessments until experience with more traditional assessments has been obtained. An intermediate step would be to include a series of separate, small-scale studies involving performance tasks to complement the other aspects of a national assessment.

Test Sophistication

Many students and assessment administrators in developing countries may be unfamiliar with some of the types of question used in the national assessment. To counteract this lack of test sophistication, detailed instructions should be written to ensure that both the student and administrator are clear about what is expected of them. Drenth (1977) suggests that in developing countries test instructions should be expanded and additional practice items added to compensate for differences attributable to lack of test sophistication.

Nonachievement Variables

Student achievement should not be interpreted in isolation from the context in which students learn. Questionnaires and rating schedules can provide valuable contextual and policy-related information and can be administered at the same time as the achievement instruments at relatively little additional expense. Contextual information might include information about teachers (for example, their qualifications and frequency of attendance at courses); class size; length of school day; teaching time; school facilities such as number and condition of desks and books; the amount of the textbook covered during the school year; time devoted to different subjects; amount of homework assigned; percentage of students being tutored outside school; and the attendance, completion, and promotion rates of students.

Identification of contextual factors related to student achievement can be particularly useful for policymakers, who can use this information to influence the reallocation of scarce financial resources. Knowledge of contextual variables can forestall policymakers' tendencies to focus on one finding or variable without considering other possible factors that might account for the finding. It can also help in the identification of manipulable variables—for example, the time allocated to curriculum

areas, the nature of pre-service and in-service teacher training, and student promotion rates—that appear to be positively related to student achievement.

Administration Manual

Because it is important that the conditions under which a national assessment is conducted be as uniform as possible from school to school, a detailed manual should be prepared by those responsible for the development of assessment instruments. Those responsible for the administration of assessments within a school should be clear about the number of students to be tested and the method of selecting the students not already preselected. Instructions should indicate how to establish appropriate conditions for assessing students and should contain a work schedule and precise details for administering tests and other instruments or tasks. A sufficient supply of materials should be available.

Review Process

Experience has underscored the necessity of pretesting all assessment instruments, including the administration manual. Test items that appear perfectly sound to the test development team are frequently confusing to students because of the wording of an item, the layout, the illustrations (when used), or the scope of the task required of the student (see box 4.3). Selection of final items or tasks usually requires some knowledge of basic psychometrics, to help, among other things, improve the reliability or internal consistency of the test or to enhance the quality of the distractors or wrong answers in multiple-choice items. Pretesting also helps provide estimates of the time needed to take the test and identifies inadequate instructions for test administrators.

Both the steering committee and the implementing agency should review tests, questionnaires, other instruments, and administrative procedures to ensure that the basic objectives of the national assessment are being addressed. Care should be taken to exclude typographical errors and inappropriate wording. All materials, including assessment instruments, questionnaires, artwork, and rating schedules, should be screened to ensure that they do not place any particular group at a disadvantage. For example, in the United States all items used in the NAEP are scrutinized to ensure that they do not unduly favor students from any particular race, culture, gender, or region.

In the review process it is important to check that test items provide an adequate and balanced sampling of the objectives of the national

Box 4.3. Dangers of Cultural Bias in Testing

The following problem was given to a sample of adolescents and adults as part of a pretest in a study of basic learning needs in Bangladesh (Greaney, Khandker, and Alam forthcoming):

> Abu owned thirty-two bighas of land. When he died his land was divided evenly among his wife, daughter, and two sons. How much land did his daughter get?

Succession and property rights under the laws of the two main religious groups that apply in Bangladesh would prompt respondents to offer answers other than eight bighas. Under Islamic law, the mother gets one-eighth of the property and each daughter gets half as much as each son. Under Hindi law, the daughter receives nothing if there are sons in the family.

curriculum. The reviewer must be aware, however, that a perfect relationship between what appears in curriculum documents and what goes on in classrooms is unlikely. When a detailed national curriculum does not exist, reviewers will rely on other sources, including the content of textbooks and their knowledge of what is covered in classrooms.

As a final stage in the review process, the steering committee and the implementing agency might develop a mock-up version of the final report that lists chapter and table headings. This helps ensure that the key objectives of the assessment are being addressed.

Administration

The logistics of administering a national assessment are substantial. They include corresponding with targeted schools to secure their cooperation; supervising the printing, packaging, and distribution of materials; recruiting and training those who will administer the assessment; organizing supervisory visits to assessment centers; and collecting, cleaning, matching, and scoring answer sheets, scripts, and questionnaires. All these activities require personnel with administrative, technical, and financial management skills. In many developing countries inadequate mail and telephone services may require local ministry of education officials to visit schools to alert them to the date and format of the assessment and to elicit their support.

Entrusting as much as possible of the actual test administration to teachers in the schools in which the assessments are being conducted will reduce administrative costs substantially. Teacher involvement can also contribute to the assessment's political viability and increase the

probability that reforms prompted by the assessment will be acted on. However, there is also a down side to the use of teachers. All teachers may not follow administration procedures adequately, giving rise to problems in comparability. There is also a danger that teachers may influence the performance of their students in a way that invalidates the assessment. When teachers are used to administer tests, it is not unusual to have a system of quality control to monitor procedures in a sample of schools.

Where there is a serious concern that the validity of an assessment may be compromised by assigning test administration to teachers, alternative strategies should be adopted. Use of the ministry of education's inspectorate, curriculum, and advisory staff to administer tests and other instruments, although not devoid of problems, may be a viable alternative. In addition to helping ensure uniformity in administration, the involvement of ministry personnel can confer a certain status on the exercise. The costs of this strategy, however, should be given serious consideration.

Analysis

Scoring, recording, data entry, cleaning, and establishment of a data base follow data collection. A data management system should be in place to monitor the quality of these activities. Here it might be possible to capitalize on the experience of public examination agencies, which have established procedures to minimize human error in conducting assessments.

Policy priorities should determine the kinds of studies undertaken. Analyses may be designed to compare group performance by location, gender, type of school attended, or ethnicity. Analyses may also be required that provide information on relationships between student achievement and the characteristics of students, schools, and teachers over time. In the United States the NAEP provides an example of analyses that consider both differences among ethnic groups and changes over time. Because the NAEP provides this kind of information, it is possible to say that in 1990, for reading at age 17, the average proficiency of white Americans was twenty-nine points higher than that of Afro-Americans. Trend data indicate that this differential is down from a fifty-three-point gap in 1971 (Elliot 1993).

The analyst should also advise policymakers on the interpretation, limitations, and implications of the results for such factors as language of instruction, level of teacher training, home background, class size, school facilities, and type of school management. When data other than

achievement data are collected, analyses should identify relationships between achievement and these factors. The results of such studies can help prevent people from arriving at simplistic conclusions—that, for example, private schools are "better" than public schools in mathematics achievement when most of the difference can be accounted for by differences in students' home backgrounds.

Results in the form of league tables, in which the test scores of schools, regions, or groups (racial or religious groups, for example) are ranked, often appeal to policymakers. However, such results should be used with considerable caution because they can be interpreted simplistically to imply causation when it is not warranted. For example, in comparing student performances in urban and rural schools, consideration has to be given to the adverse circumstances in which some schools operate. The added achievement value in schools with low test scores may actually be greater than in schools with higher test scores.

Reporting

Assessment results should be reported as soon as possible after data collection. Lengthy delays diminish the usefulness of the exercise. The report should be concise, simply written, and devoid of educational jargon (Shepard 1980). It should feature simple graphs and bar charts. Harried policymakers are badly served when presented with volumes of cross-tabulations to wade through to discover key findings. The timely, well-presented, and well-illustrated reports produced by the NAEP, and recently by the IEA, can serve as models.

Conclusions should always be based on unambiguous evidence derived from the data, and the report should document relevant procedures and criteria so that readers can assess the significance of conclusions. The preparation of press releases by the implementing agency can help reduce the possibility of misinterpretation by the media. Oral reports may be provided for such key interest groups as senior ministry of education officials, inspectors, teacher training authorities, teachers' and school managers' representatives, curriculum authorities, and textbook publishers. Seminars can also be organized for teachers to discuss the results and their implications.

Many approaches have been used in national and international assessments in reporting results. One involves reporting average levels of student performance in a curriculum area. The others involve reporting the percentage of students associated with specified achievements. The achievements, however, are defined in different ways.

Average Performance of Students in a Curriculum Area

If the individual scores of a representative sample of students in a country are added and then divided by the number of students, one gets an overall average for performance in a particular curriculum area, at a particular age, for that country. The procedure may not be quite as simple as this in practice, since adjustments may have to be made to take into account disproportional sampling of students in different regions or types of school. The basic point, however, is that one is seeking to represent in quantitative terms the average level of performance in the country.

This information is of limited value in itself because it does not tell us whether the average obtained can be regarded as "satisfactory" or "unsatisfactory." It can be useful, however, if comparative data are available with which the obtained average score can be compared. Thus, the information could be useful as an indication of whether standards in the country were, in general, stable, rising, or falling, if comparable information were available from an earlier point in time. It could also be useful if similar information were available from other countries, as is the case in international studies of assessment. Both the IEA and the IAEP have reported mean scores for participating countries in a variety of curriculum areas. The Organisation for Economic Co-operation and Development (OECD) has made use of these data to highlight differences in achievement among its member countries (OECD 1995).

Percentage Passing Items

Some national (for example, the U.S. NAEP) and international (for example, the IAEP) assessments have reported results at the individual-item level. For each individual item, the percentage of students answering correctly was reported. Average percent-correct statistics were used to summarize the results (see Baker and Linn 1995; Phillips and others 1993). This information is probably too detailed for most readers. Furthermore, if comparisons are to be made from one assessment to another or among the results for different grades, this approach requires that identical sets of items be used.

Percentage Achieving Mastery of Curriculum Objectives

In another approach, the percentages of students who achieve mastery of main curriculum objectives are presented. In one Irish assessment the mathematics curriculum for students in grades 5 and 6 was divided into

fifty-five separate objectives in computation, concepts, and problem solving. Objectives called, for example, for the student to be able to:

- Add a column of numbers containing not more than five digits
- Subtract two numbers containing not more than five digits
- Perform simple arithmetic operations involving zero
- Identify common factors between two numbers.

A student was regarded as having mastered an objective when he or she correctly answered a specified number of items per objective on a multiple-choice written test. Statistics were provided for each of the fifty-five objectives, indicating the percentage of students who had mastered the objective. Aspects of the national curriculum that posed problems were identified. Objectives students had trouble solving included:

- Dividing a fraction by a whole number and vice versa
- Solving algebraic equations that call for two simple arithmetical operations
- Identifying the least common multiple of two numbers (Kellaghan and others 1976).

Percentage Achieving Specified Attainment Targets

In some education systems specific attainment targets are set for students at varying points in their educational careers. Where this is the case, an assessment system may be designed to obtain estimates of the number of students who are reaching these targets. In the British system the extent to which students are meeting attainment targets of the national curriculum at ages 7, 11, 14, and 16 is identified. Each target is divided into levels of ascending difficulty on a scale of 1 to 10, with clear criteria defining what a student must know, understand, or be able to do to be rated as scoring at that level. The assessments of 7-year-olds concentrate on levels 1 to 3. There were thirty-two targets relevant to 7-year-olds in the 1991 assessment: five in English, thirteen in mathematics, and fourteen in science. Results were reported as the percentages of students who satisfied each level in each curriculum area.

In English, the percentages attaining levels 1, 2, and 3 were given for the five targets: speaking and listening, reading, writing for meaning, spelling, and handwriting. In mathematics, examples of targets for which results were presented were number, algebra, and measures; using and applying mathematics; number and number notation; number operations (+, −, ÷, ×); and shape and space (two- and three-dimensional shapes). The science targets included life processes, genetics and evolu-

tion, human influences on the Earth, types and uses of materials, energy, and sound and music (Great Britain, Department of Education and Science, 1991).

Percentage Functioning at Specified Levels of Proficiency

Another way of presenting results, used in several state and national assessments, is to construct a proficiency scale through statistical procedures and determine levels on the scale through judgmental processes. This, as we saw in chapter 2, was done in the NAEP. It was also done in the Canadian national assessment of 13- and 16-year-olds in mathematics, science, reading, and writing, for which five proficiency levels were established in each curriculum area (Canada, Council of Ministers of Education, 1996). Each level is described in terms of the knowledge and skills that a student operating at the level should exhibit; the percentage of students functioning at each level is then reported. In science the student should be able to describe, at a given level:

Level 1	Physical properties of objects
Level 2	Qualitative changes in the properties of a substance when it is heated or cooled
Level 3	The structure of matter in terms of particles
Level 4	Qualitatively, a chemical reaction or phase change
Level 5	Quantitatively, the product of a reaction given the reactants, or vice versa.

Sometimes labels are attached to levels. For example, students in the NAEP are described as lacking basic competency, as having attained basic competency, as being proficient, or as being advanced. An alternative nomenclature was used in an assessment in Kentucky: students were described as novice, apprentice, proficient, or distinguished (Guskey 1994). Although such labels have some attractive features, they can have negative connotations. They can also mean different things at different grade levels, and even for the same grade they are likely to be interpreted in different ways by different people. When results are reported as levels, it would seem preferable not to go beyond the verbal description of what a student at the level knows or can do.

Cost-Effectiveness

One of the arguments in favor of holding national assessments is that they can lead to economic efficiencies in the education system. The results of such assessments can, for example, pinpoint areas of the

Table 4.2. Distribution of Costs of Components of National Assessment in the United States

Component	*Percentage of total cost*
Instrument development	15
Sampling and selection	10
Data collection	30
Data processing	10
Data analysis	15
Reporting and dissemination	15
Governance	5

Source: Koeffler 1991.

curriculum in which students are achieving well or poorly, thereby prompting curriculum revision. The linking of teacher questionnaire and achievement data can identify teacher needs and provide the content for in-service courses.

Cost is an important consideration in deciding whether to initiate a national assessment. Against a background of competing demands for scarce resources, costs have to be justified and kept as low as possible. The data produced should be of sufficient value to justify expenditure. Unless initial cost estimates are within budget limits, the original design for the assessment will have to be modified (Ilon 1996).

Test and questionnaire development and other technical components of a national assessment may account for a relatively small percentage of the overall budget. Experience suggests that other elements of an assessment prove much more expensive. In Chile, for example, the technical components accounted for 10 percent of the budget; the remaining 90 percent was spent on such matters as printing, distribution, data gathering, data processing, and distribution of the report (Himmel 1996). Costs in the United States for data collection, processing, analysis, and reporting were considerably larger than those for instrument development and sampling (table 4.2; Koeffler 1991).

Loxley's advice to set aside a contingency fund for emergencies, although intended for those involved in international assessment, is also relevant to national assessment. In recommending that 10 percent of the budget be earmarked for this purpose, he notes that "it is never a question of whether emergencies will arise, but rather of when and how many" (1992, p. 293).

Conclusion

A well-designed and well-executed national assessment can provide a country with useful indicators of the health of its education system. Given

the percentage of central government expenditure devoted to education (more than 10 percent in many developing countries), the need to monitor this investment scarcely needs to be justified. However, high levels of technical competence and of administrative and political skill are required to conduct an effective assessment. The appendix to this book presents a checklist of the components of a model assessment for use by practitioners.

To assume that the results of even a well-administered national assessment will inevitably bring about immediate (and positive) change in an education system would be naive. Any proposed reform must confront the potentially conflicting interests and values of the different stakeholders in the education system. Stakeholders include religious, ethnic, and language groups; teachers' unions or organizations; representatives of school management; parents; school inspectors; curriculum specialists; and textbook publishers. National assessment data on the functioning of a school system can play an important role in ensuring that the selection of educational priorities does not depend solely on the values of politically powerful groups.

5
Pitfalls of National Assessment: A Case Study

Guidelines for the conduct of a national assessment were given in chapter 4. We present here a fictitious case study of a national assessment to allow readers to review their understanding of the guidelines. The reader's task is to detect examples of poor practice in the assessment. The case is analyzed at the end of the chapter, and some of the more obvious examples of poor practice are identified and discussed.

Background to the Initiation of a National Assessment in Sentz

During the annual meeting of the National Development Authority of Sentz, chaired by the prime minister, industrialists and businesspeople complained that in their view standards of education throughout the country had fallen sharply since independence. Numerous examples of workers' inability to do basic computations, solve simple mathematical problems, or communicate effectively either orally or in writing were cited. Two weeks later, the National Chamber of Commerce complained that the failure of the secondary school system to develop basic skills posed a serious threat to the country's future economic development. Not unexpectedly, these charges received prominent attention in feature articles and in editorials in the national and regional news media. Sensitive to the charges leveled against his government, the prime minister informed parliament that he was ordering the Ministry of Education to conduct a review of educational standards in the principal secondary school subjects.

School System

Sentz is a large low-income country that gained independence in 1968. Most of its population resides in isolated mountain villages. The country offers six years of free noncompulsory primary school education and

five years of fee-paying secondary education. Following independence, the government undertook a major program of investment in education. The number of schools and the size of the student body have expanded rapidly since 1970; other data indicate a decline in quality of education (table 5.1).

In 1969 Kubsa replaced English as the official language. Kubsa is the mother tongue of 30 percent of the population; an additional 40 percent speak it as a second language. It is also the medium of instruction in schools. The Ministry of Education has been charged by Parliament with the task of promoting Kubsa. Other principal languages are Brio, spoken by 45 percent of the population and Holog, spoken by 15 percent. The remaining 10 percent of the people living in remote areas speak other languages. In secondary school the compulsory subjects are Kubsa, mathematics, English, religious education, science, and agriculture.

Response to Education Concerns

The minister for education responded to the prime minister's initiative by announcing a reform program "to improve the quality of student learning in the key subject areas." He reiterated the prime minister's concern over the lack of problem-solving skills in mathematics and the poor levels of oral fluency and writing skills in both Kubsa and English. As a first stage in the reform, he promised a rigorous scientific analysis of existing standards in secondary schools. "To help guide instruction, we need information on what students know and what they do not know," he added. The information could be used in the future to guide curriculum reform. Baseline data would be established to monitor changes in achievement standards over time. He indicated that the ten

Table 5.1. Educational Developments in Sentz, 1970–90

Item	*1970*	*1980*	*1990*
Number of secondary schools	164	176	236
Number of students	26,841	121,637	162,249
Female students as percentage of total	27	38	41
Gross enrollment ratio[a]	4	9	11
Student-teacher ratio	28	44	45
Expenditure per student as a percentage of GNP per capita	104	67	65
Number of textbooks per student	2.2	1.2	1.1
Percentage of teachers trained	75	54	57

a. The number of secondary students expressed as a percentage of the total population of individuals of secondary school age.

highest-scoring schools in the assessment would be acknowledged formally by the minister. Low-scoring schools would be visited by inspectors with the object of improving standards. Data were to be collected on a regional basis so that the ministry could introduce programs in low-scoring regions. A budget equivalent of US$25,000 was to be set apart from the current education budget to finance the assessment.

The Association of Secondary Teachers opposed the national assessment. The association claimed in a press release that the exercise reflected on the integrity of teachers, that it was an unsubtle attempt to hold teachers accountable for their students' achievements, and that it failed to take into consideration the different economic and social circumstances under which schools operated. The association also claimed that a shortage of textbooks and the inadequacy of the in-service program following curriculum changes were primarily responsible for the perceived drop in standards. It concluded that it would be naive of the ministry to expect teacher cooperation in the conduct of the national assessment until a two-year-old promise to increase teacher salaries had been honored.

National Assessment of Educational Standards in Sentz

Organization

A *National Committee* was established to oversee the conduct of the national assessment. It consisted of three senior inspectors from the secondary section of the ministry, two professors of education from the national university, the director of the National Curriculum Authority, the director of the National Examinations Board, and two prominent businesspeople. Following its first meeting, the National Committee invited proposals for the design and conduct of the assessment. Three submissions were received. One was drafted by a research unit within the university, a second came from a private consultancy group, and a third came from the planning section of the Ministry of Education. The contract was awarded to the planning section, which had submitted the lowest bid and would, it was felt, have less difficulty in gaining access to schools than other agencies.

A special *Working Group* was established within the ministry to develop tests, sample schools, and arrange for administration, scoring, analysis, and reporting. Two middle-rank staff members—a higher executive officer (with a master's degree in the philosophy of education) and a staff assistant—were appointed on a full-time basis. The officer

was to serve as chief executive officer of the Working Group and as an ex officio member of the National Committee. Other staff were to be made available by the National Examinations Board as needed. Within the ministry it was agreed that the secondary school inspectorate would administer the tests.

Test Development

Design. The Working Group submitted its plan to the National Committee to assess students in grades 1, 3, and 5 (the final year of secondary school) at the end of the school year. Subjects for study were Kubsa, mathematics, English, and science. Testing was scheduled for early June to avoid a clash with the all-important school leaving examination, which was scheduled for the last two weeks of that month. Multiple-choice tests were to be developed for each of the four subject areas. It was planned to administer the tests every two years.

Test content. Two language tests, to be developed locally, would contain subtests in vocabulary, paragraph comprehension, spelling, and grammar. The science test, also to be constructed locally, would contain separate sections on the Earth and the atmosphere, forces, electricity and magnetism, and energy. An internationally known, commercially produced mathematics test was to be translated into Kubsa; the technical manual for the mathematics test indicated that it had "adequate content validity" and could be considered highly reliable, with an internal consistency index of 0.91.

Item writers. To preserve the security of the tests and to reduce costs, former school inspectors and retired university professors would be recruited to write items for the three locally produced tests. Their subject-area knowledge would help ensure that the tests would be accepted as valid.

Sample. The Working Group recommended that the assessment be confined to six regions because dependable data on student and school numbers were not available for the remaining four, more remote, regions. The group accepted that results from the sampled regions would probably represent an overestimate of national achievement levels in the four targeted subjects. A visiting sampling expert advised that the six regions targeted be pooled and that a random sample of schools be drawn. Sample size would be determined by the need to have the mean for the population within 3 percent of the estimated mean with a 95 percent confidence level. Within each selected school every second student (randomly selected) would be assessed.

Implementation

Schedule. A schedule of activities to be followed during the course of the national assessment was developed (table 5.2).

Publication of results. The Working Group recommended that the normal departmental convention on what information should be released to the public be followed and that decisions about publication rest with the minister for education.

Modifications. The National Committee accepted the Working Group's proposals. The committee suggested, however, that the scope of the assessment was unduly ambitious, given staffing and budgetary constraints, and recommended that the grade 3 sample and the science component be dropped. In defense of this recommendation, the committee

Table 5.2. Schedule of Activities for a National Assessment in Sentz

Activity	*Responsibility*	*Deadline*
1 Establish National Steering Committee	Ministry of Education	Early December
2 Determine terms of reference	National Steering Committee	Late December
3 Appoint implementing agency	National Steering Committee	March 1
4 Define precise objectives	Implementing agency	May 1
5 Review relevant studies and data	Implementing agency	May 1
6 Design and select sample	Implementing agency	July 1
7 Develop instruments and manuals	Implementing agency	May 1–August 31
8 Pretest instruments and manuals	Implementing agency	September 1–October 31
9 Print final forms of instruments; determine time limits; train administrators	Implementing agency	November 1–December 31
10 Administer instruments	Implementing agency	January
11 Create data base	Implementing agency	February–March
12 Analyze data	Implementing agency	April–June
13 Draft report	Implementing agency	July–October
14 Discuss drafts	Implementing agency	November–February
15 Prepare final report	Implementing agency	March–April
16 Disseminate to maximize impact	Implementing agency	May onward

argued that given the high quality of the science curriculum and of the national school leaving examination, issues of standards in science could be addressed through an analysis of public examination results. Finally, the committee proposed that the final report should be brief, devoid of technical language, and limited to a detailed presentation of the results and their implications for secondary education in Sentz.

Analysis of the Case

The inspiration for the initiation of a national assessment in Sentz was not unusual, but the groundwork was insufficient to validate the claim that standards had fallen. At the outset, consideration should have been given to the extent to which the school-going population had changed over the intervening years. The data (table 5.1) support the notion that the secondary school population had indeed changed. Key changes noted include the rapid increase in student numbers, a steady increase in female participation, and an almost threefold increase in the gross enrollment rate. Such changes could explain perceived failures in the school system and invalidate the charge that the school system had not helped children develop basic skills. However, the data also point to a sharp deterioration in student-teacher ratios, in public expenditure on secondary education, in textbook provision, and in levels of teacher qualification.

Responses to Assessment

The minister's generally reasonable response suggests that national assessment is to become a feature of the education system. Although the proposed study will not show what students need to know, it can provide useful information on what students know at present and baseline data for long-term monitoring of standards.

There is a danger that the proposed assessment will become a high-stakes exercise if individual teachers feel that their work is being monitored by an external agency and that their schools are being ranked for close scrutiny. They may decide not to participate in the assessment. Some may resort to strategies such as teaching to the test that would render findings suspect. Early communication with the teachers' union or inclusion of the union in the National Committee might have weakened or eliminated opposition to the assessment. If teachers remain strongly opposed, the exercise may have to be abandoned.

Implementation Procedures

Organization of the assessment. The membership of the National Committee needs to be justified. What comparative advantage do business personnel have over other possible representatives? Does the committee have any technical competence? Whose interests are best served by the committee? How might its perspective be broadened? Because curriculum reform seems to be a priority, should textbook publishers be represented? Are there good political reasons for excluding representatives of the teaching profession?

The dependability of ministry officials in commitment, quality of work, and ability to honor deadlines was not taken into account. Teachers may perceive the assessment as a high-stakes exercise because of the presence of ministry personnel such as school inspectors in schools. Other issues to consider include whether the ministry should conduct an evaluation of its own services and whether it will be willing to release information that might be considered politically embarrassing.

The technical competence of the Working Group members is a crucial issue. For example, a person with a master's degree in philosophy of education will not have acquired the necessary competencies in measurement and research as part of his or her training. Seniority is a decided asset and can give an assessment the necessary status within the ministry. The support of other staff "as needed" may be an unsatisfactory arrangement; key personnel may have prior obligations to the public examination system and may not be available when needed. The issue of payment for additional help was not addressed.

Costs. Estimates of such costs as travel, subsistence, test development, use of computers, analyses, and writing and publishing of reports are not well substantiated. Experience in some developing countries suggests that items such as travel, subsistence, printing, and supplies tend to be underbudgeted. In addition, the true costs of the assessment to the ministry are not being considered in the contract. For example, if tests and instructions are printed by a government printing office out of the ministry's own resources, it is evident that the budget does not reflect the true cost. Thus, other bidders for the contract were placed at a considerable disadvantage. In assessing costs, the ministry and the planning section need to consider the salaries of all staff members who participate in the assessment. Are salaries to be paid from the project budget or absorbed by the ministry? If the ministry absorbs salary costs, the opportunity costs need to be calculated.

Design. The plan is overambitious. Individual subject areas need not be assessed as frequently as every two years; improvements or declines in achievement levels tend to be more gradual. An assessment of an

individual curricular area once every five years would be adequate. In many industrial countries, national assessments are confined to a much smaller number of subjects and grade levels. It will not be possible to address the prime minister's concerns about inadequate oral and written skills through the use of multiple-choice tests. Testing any secondary grade, especially grade 5, is not recommended immediately prior to "all-important" school leaving examinations.

There is no evidence that a model of the educational process guided the national assessment design. But a model is important for discriminating among the many aspects of the educational process that will be selected for study and for interpreting findings. No reference is made to the context in which learning takes place; the design does not include key variables that might help explain differences in student achievement (for example, school and class size, attendance, resources such as books, trained teachers, and student background characteristics).

Test content. The "adequate content validity" claim for the commercial mathematics test is not sufficient justification for using it. Appropriateness has to be established for students in Sentz.

Language. Children with Kubsa as a first language—a minority of the students—will have an advantage over others in the national assessment.

Item writers. University professors may have little technical competence in test construction or little insight into the actual curriculum pursued in schools. Their subject mastery may not be adequate if it does not include the precise subject matter for the targeted grades. Practicing teachers, given some training in test development techniques, may be more effective.

Sample. Given the limitations of the national data base, the approach and assumptions made by the Working Group seem defensible. The emphasis on the margin of error is correct and helps avoid unwarranted assumptions being made about groups of interest who record different scores. Randomly selecting students within classes can be disruptive and unnecessary. Professional advice on sampling issues is highly recommended.

Schedule. At first glance the proposed schedule seems unnecessarily drawn out. The timetable presented for the case study is based on one developed by experienced national evaluators for a national assessment in a developing country, and some item-writing had already been completed. (In the case study used as a model by the Sentz planners, it was felt that technical assistance would be required for activities 6, 11, 12, and 13; see table 5.2). Although the natural enthusiasm of a minister for quick results should be resisted if it involves seriously compromising the design and implementation of the assessment, there is a danger that

by the time results are available, after a long study, political interest in the project may have waned and the probability of institutionalizing national assessment diminished. When possible, dates for the administration of tests and questionnaires should be the same throughout the country. If test dates vary, items may be leaked to some schools in advance of the tests.

Publication of results. The sensitive question of publication and editorial rights should be resolved at the beginning. An open approach can lead to cooperation by potential key stakeholders. In some countries confidence in results edited by the ministry of education may be limited. The ministry in Sentz, for instance, might find it politically difficult to provide evidence that a particular policy, such as achieving equality of student achievement in different geographic areas, was unsuccessful.

Modifications. The proposals made by the Working Group appear realistic given the circumstances. The use of public examination results to monitor standards would, however, be of little value. The emphasis on a brief report, devoid of technical language, is good. The purposes and procedures of the national assessment should be described clearly and in sufficient detail so that interested parties in the country can understand and debate the issues and the findings.

A Choice to Make

The conduct of a national assessment requires compromise and ingenuity—a perfectly planned and executed study is not achievable. But even when a national assessment is well planned and well executed, it may be argued that spending money on this kind of study is not justified when the system lacks resources for new schools and textbooks. The resources required for the conduct of a national assessment, however, would not go far in addressing significant shortcomings in school facilities or textbooks, and the information obtained through a national assessment can bring about cost-efficiencies by identifying failing features of existing arrangements or by producing evidence to support more effective alternatives. It is up to the proponents of a national assessment to show that the likely benefits to the education system as a whole merit the allocation of the necessary funds.

References

Allen, R., N. Bettis, D. Kurfman, W. MacDonald, Ina V. S. Mullis, and C. Salter. 1990. *The Geography Learning of High-School Seniors.* Princeton, N.J.: National Assessment of Educational Progress, Educational Testing Service.

Anderson, Lorin W., and T. Neville Postlethwaite. 1989. "What IEA Studies Say about Teachers and Teaching." In Alan C. Purves, ed., *International Comparisons and Educational Reform.* Washington, D.C.: Association for Supervision and Curriculum Development.

Anderson, Lorin W., Doris W. Ryan, and Bernard J. Shapiro. 1989. *The IEA Classroom Environment Study.* Oxford: Pergamon.

Anderson, Lorin, L. B. Jenkins, J. Leming, W. MacDonald, Ina V. S. Mullis, M. Turner, and J. S. Wooster. 1990. *The Civics Report Card: Trends in Achievement from 1976 to 1988 at Ages 13 and 17; Achievement in 1988 at Grades 4, 8, and 12.* Princeton, N.J.: National Assessment of Educational Progress, Educational Testing Service.

Applebee, Arthur N., Judith A. Langer, Lynn B. Jenkins, Ina V. S. Mullis, and M. A. Foertsch. 1990a. *Learning to Write in Our Nation's Schools: Instruction and Achievement in 1988 at Grades 4, 8, and 12.* Princeton, N.J.: National Assessment of Educational Progress, Educational Testing Service.

Applebee, Arthur N., Judith A. Langer, Ina V. S. Mullis, and Lynn B. Jenkins. 1990b. *The Writing Report Card, 1984 to 1988: Findings from the Nation's Report Card.* Princeton, N.J.: National Assessment of Educational Progress, Educational Testing Service.

Baker, Eva L., and Robert L. Linn. 1995. "United States." In *Performance Standards in Education: In Search of Quality.* Paris: Organisation for Economic Co-operation and Development.

Bali, S. K., Peter J. D. Drenth, Henk van der Flier, and W. C. Young. 1984. *Contribution of Aptitude Tests to the Prediction of School Performance in Kenya: A Longitudinal Study.* Lisse, The Netherlands: Swetts and Zeitlinger.

Báthory, Zoltán. 1989. "How Two Educational Systems Learned from Comparative Studies: The Hungarian Experience." In Alan C. Purves, ed., *International Comparisons and Educational Reform.* Washington, D.C.: Association for Supervision and Curriculum Development.

Bennett, Neville, and Charles Desforges. 1991 "Primary Education in England: A System in Transition." *Elementary School Journal* 92:61–78.

Bloom, Benjamin S., George F. Madaus, and Thomas Hastings. 1981. *Evaluation to Improve Student Learning.* New York: McGraw-Hill.

Bottani, Norberto. 1990. "The Background of the CERI/OECD Project on International Educational Indicators." *International Journal of Educational Research* 14:335–42.

Bottani, Norberto, and Albert Tuijnman. 1994. "International Education Indicators: Framework, Development and Interpretation." In *Making Education Count. Developing and Using International Indicators.* Paris: Organisation for Economic Co-operation and Development.

Bottani, Norberto, and Herbert W. Walberg. 1994. "International Educational

Indicators." In Torsten Husén and T. Neville Postlethwaite, eds., *The International Encyclopedia of Education*, Vol. 5, 2d ed. Oxford: Pergamon

Broadfoot, Patricia, D. Abbot, P. Croll, M. Osborn, and A. Pollard. N.d. "Look Back in Anger. Findings of the PACE Project Concerning Primary Teachers' Experiences of SATs." University of Bristol, U.K.

Buber, Martin. 1963. *Israel and the World: Essays in a Time of Crisis*. New York: Schocken Books.

Burnstein, Leigh, Jeannie Oakes, and Gretchen Guiton. 1992. "Education Indicators." In Marvin C. Alkin, ed., *Encyclopedia of Educational Research*, Vol. 2. 6th ed. New York: Macmillan.

Canada, Council of Ministers of Education. 1996. *School Achievement Indicators Program. Science Assessment. Framework and Criteria*. Toronto, Ontario.

Carroll, John B. 1975. *The Teaching of French as a Foreign Language in Eight Countries. International Studies in Evaluation*, Vol. 5. Stockholm: Almquist and Wiksell.

Chinapah, Vinayagum. 1992. "Monitoring and Surveying Learning Achievements: A Status Report." Studies and Working Documents 1. Paris: UNESCO.

Cizek, Gregory J. 1991. "Innovation or Enervation? Performance Assessment in Perspective." *Phi Delta Kappan* 72:695–99.

Cohen, Michael. 1987. "Improving School Effectiveness: Lessons from Research." In Virginia Richardson-Koehler, ed., *Educators' Handbook. A Research Perspective*. New York: Longman.

Comber, L. C., and John P. Keeves. 1973. *Science Education in Nineteen Countries: An Empirical Study: International Studies in Evaluation*, Vol. 1. Stockholm: Almquist and Wiksell.

Darling-Hammond, Linda, and A. Lieberman. 1992. "The Shortcomings of Standardized Tests." *Chronicle of Higher Education* 39 January 29, B1–B2.

Dearing, Ron. 1993. *The National Curriculum and Its Assessment. Final Report*. York, U.K.: National Curriculum Council, and London: School Examinations and Assessment Council.

Dedze, Indra. 1995. "Reading Achievement within the Educational System of Latvia: Results from the IEA Reading Literacy Study." Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Dossey, J. A., Ina V. S. Mullis, M. M. Lindquist, and D. L. Chambers. 1988. *The Mathematics Report Card: Are We Measuring Up? Trends and Achievement Based on the 1986 National Assessment*. Princeton, N.J.: National Assessment of Educational Progress, Educational Testing Service.

Drenth, Peter J. D. 1977. "Prediction of School Performance in Developing Countries. School Grades or Psychological Tests?" *Journal of Cross-Cultural Psychology* 8:49–70.

Drenth, Peter J. D., Henk van der Flier, and Issa M. Omari. 1983. "Educational Selection in Tanzania." *Evaluation in Education* 7:93–217.

Elley, Warwick B. 1992. *How in the World Do Students Read? IEA Study of Reading Literacy*. The Hague: International Association for the Evaluation of Educational Achievement.

_____. 1994. *The IEA Study of Reading Literacy : Achievement and Instruction in Thirty-two School Systems*. Oxford: Pergamon.

Elliot, E. J. 1993. "National Testing and Assessment Strategies: Equity Implications of Leading Proposals for National Assessments." Paper presented at the Equity and Educational Testing Assessment Seminar, Washington, D.C., March 12.

Fitz-Gibbon, C. T. 1995. "Indicators for Quality in the UK: Examinations and the OFSTED System." Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Foxman, D. D., G. Ruddock, and I. McCallum. 1990. *Assessment Matters: No. 3 APU Mathematics Monitoring 1984–1988, Phase 2.* London: School Examination and Assessment Council.

Foxman, D., D. Hutchinson, and B. Bloomfield. 1991. *The APU Experience 1977–1990.* London: School Examination and Assessment Council.

Foxman, D. D., M. J. Creswell, M. Ward, M. E. Badger, J. A. Tuson, and B. A. Bloomfield. 1980a. *Mathematical Development, Primary Survey Report No. 1.* London: Her Majesty's Stationery Office.

Foxman, D. D., R. M. Martini, J. A. Tuson, and M. J. Creswell. 1980b. *Mathematical Development, Secondary Survey Report No. 1.* London: Her Majesty's Stationery Office.

Frith, D. S., and H. G. Macintosh. 1984. *A Teacher's Guide to Assessment.* Cheltenham: International Association for Educational Assessment.

Gipps, Caroline. 1993. "National Curriculum Assessment in England and Wales." Paper presented at the ICRA Conference, University of London Institute of Education.

Gipps, Caroline, and Harvey Goldstein. 1983. *Monitoring Children. An Evaluation of the Assessment of Performance Unit.* London: Heinemann.

Gipps, Caroline, and Patricia Murphy. 1994. *A Fair Test? Assessment, Achievement and Equity.* Buckingham, U.K.: Open University.

Gipps, Caroline, Bet McCallum, Shelley McAllister, and Margaret Brown. 1991. "National Assessment at Seven: Some Emerging Themes." Paper presented at British Educational Research Association Annual Conference.

Gorman, Tom P., Alan Purves, and R. E. Degenhart. 1988. *The IEA Study of Written Composition 1: The International Writing Tasks and Scoring Scales.* Oxford: Pergamon.

Graves, Donald. 1983. *Writing: Teachers and Children at Work.* Portsmouth, N.H.: Heinemann.

Greaney, Vincent. 1996. "Stages in National Assessment." In Paud Murphy, Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas, eds., *National Assessments: Testing the System.* EDI Learning Resources Series. Washington, D.C.: World Bank.

Greaney, Vincent, and John Close. 1989. "Mathematics Achievement in Irish Primary Schools." *Irish Journal of Education* 22:51–64.

Greaney, Vincent, and Thomas Kellaghan. 1996. "The Integrity of Public Examinations in Developing Countries." In Harvey Goldstein and Toby Lewis, eds., *Assessment: Problems, Developments and Statistical Issues. A Volume of Expert Contributions.* Chichester, Sussex, U.K.: Wiley.

Greaney, Vincent, Shahidur Khandker, and Mahmudul Alam. Forthcoming. "Bangladesh: Assessing Basic Learning Skills." Washington, D.C.: World Bank.

Great Britain. Department of Education and Science. 1988. *National Curriculum. Task Group on Assessment and Testing. A Report.* London.

Great Britain. Department of Education and Science. 1991. *Testing 7-Year-Olds in 1991 : Results of the National Assessments in England.* London.

Guskey, Thomas R. 1994. *High Stakes Performance Assessment: Perspectives on Kentucky's Educational Reform.* 2d ed. Thousand Oaks, Calif.: Corwin Press.

Haladyna, Thomas M., Susan B. Nolan, and Nancy S. Haas. 1991. "Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution." *Educational Researcher* 20 (5):2–7.

Hammack, D. C., M. Hartoonian, J. Howe, L. B. Jenkins, L. S. Levstick, W. B. MacDonald, Ina V. S. Mullis, and Eugene Owen. 1990. *The U.S. History Report Card: The Achievement of Fourth-, Eighth-, and Twelfth-Grade Students in 1988 and Trends from 1986 to 1988 in the Factual Knowledge of High-School Juniors.* Princeton, N.J.: National Assessment of Educational Progress, Educational Testing Service.

Himmel, Erika. 1996. "National Assessment in Chile." In Paud Murphy, Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas, eds., *National Assessments: Testing the System.* EDI Learning Resources Series. Washington D.C.: World Bank.

Horn, Robin, Laurence Wolff, and Eduardo Velez. 1991. "Developing Educational Assessment Systems in Latin America: A Review of Issues and Recent Experience." Staff Working Paper 9. World Bank, Latin America and the Caribbean Technical Department, Washington, D.C.

Husén, Torsten. 1987. "Policy Impact of IEA Research." *Comparative Education Review* 31:29–46.

Husén, Torsten, ed. 1967. *International Study of Achievement in Mathematics.* 2 vols. Stockholm: Almquist and Wiksell.

Ilon, Lynn. 1996. "Considerations for Costing a System." In Paud Murphy, Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas, eds., *National Assessments: Testing the System.* EDI Learning Resources Series. Washington, D.C.: World Bank.

Instituto SER de Investigación/Fedesarrollo. 1994. "Educación." *Coyuntura Social* 11:11–24.

Izard, John. 1996. "The Design of Tests for National Assessment Purposes." In Paud Murphy, Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas, eds., *National Assessments: Testing the System.* EDI Learning Resources Series. Washington, D.C.: World Bank.

Johnson, Eugene G. 1992. "The Design of the National Assessment of Educational Progress." *Journal of Educational Measurement* 29:95–110.

Johnstone, James N. 1981. *Indicators of Education Systems.* Paris: UNESCO.

Keeves, John P. 1994. "Examinations: Public." In Torsten Husén and T. Neville Postlethwaite, eds., *The International Encyclopedia of Education,* Vol. 4. 2d ed. Oxford: Pergamon.

Keeves, John P., and Malcolm J. Rosier, eds. 1992. *The IEA Study of Science 1: Science Education and Curricula in Twenty-three Countries.* Oxford: Pergamon.

Kellaghan, Thomas. 1994. "Family and Schooling." In Torsten Husén and T. Neville Postlethwaite, eds., *The International Encyclopedia of Education,* Vol. 4. 2d ed. Oxford: Pergamon.

_____. 1996a. "Can Public Examinations Be Used to Provide Information for National Assessment?" In Paud Murphy, Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas, eds., *National Assessments: Testing the System.* EDI Learning Resources Series. Washington, D.C.: World Bank.

_____. 1996b. "IEA Studies and Educational Policy." *Assessment in Education* 3:143–60.

_____. 1996c. "National Assessment in England and Wales." In Paud Murphy,

Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas, eds., *National Assessments: Testing the System*. EDI Learning Resources Series. Washington, D.C.: World Bank.

Kellaghan, Thomas, and George F. Madaus. 1982. "Educational Standards in Great Britain and Ireland." In Gilbert R. Austin and Herbert Garber, eds., *The Rise and Fall of National Test Scores*. New York: Academic Press.

Kellaghan, Thomas, and Vincent Greaney. 1992. *Using Examinations to Improve Education. A Study in Fourteen African Countries*. World Bank Technical Paper 165. Washington, D.C.

Kellaghan, Thomas, and Aletta Grisay. 1995. "International Comparisons of Student Achievement: Problems and Prospects." In *Measuring What Students Learn. Mesurer les Résultats Scolaires*. Paris: Organisation for Economic Co-operation and Development.

Kellaghan, Thomas, George F. Madaus, Peter W. Airasian, and Patricia J. Fontes. 1976. "The Mathematical Attainments of Post-Primary School Entrants." *Irish Journal of Education* 10:3–17.

Kellaghan, Thomas, Kathryn Sloane, Benjamin Alvarez, and Benjamin S. Bloom. 1993. *The Home Environment and School Learning. Promoting Parental Involvement in the Education of Children*. San Francisco: Jossey Bass.

Kifer, Edward. 1989. "What IEA Studies Say about Curriculum and School Organizations." In Alan C. Purves, ed., *International Comparisons and Educational Reform*. Washington, D.C.: Association for Supervision and Curriculum Development.

Koeffler, S. 1991. "Assessment Design." Paper presented at the Seminar on Measurement/Assessment Issues, Educational Testing Service, Princeton, N.J.

Lambin, Rosine. 1995. "What Can Planners Expect from International Quantitative Studies?" In Wilfred Bos and Rainer M. Lehmann, eds., *Reflections on Educational Achievement. Papers in Honour of T. Neville Postlethwaite*. New York: Waxman.

Langer, Judith A., Arthur N. Applebee, Ina V. S. Mullis, and M. A. Foertsch. 1990. *Learning to Read in Our Nation's Schools: Instruction and Achievement in 1988 at Grades 4, 8, and 12*. Princeton, N.J.: National Assessment of Educational Progress, Educational Testing Service.

Lapointe, Archie, Nancy A. Mead, and Janice M. Askew. 1992. *Learning Mathematics*. Princeton, N.J.: Educational Testing Service.

Lapointe, Archie E., Janice M. Askew, and Nancy A. Mead. 1992. *Learning Science*. Princeton, N.J.: Educational Testing Service.

Lapointe, Archie E., Nancy A. Mead, and Gary W. Phillips. 1989. *A World of Differences: An International Assessment of Mathematics and Science*. Princeton, N.J.: Educational Testing Service.

Le Mahieu, Paul G. 1984. "The Effects on Achievement and Instructional Content of a Program of Student Monitoring through Frequent Testing." *Educational Evaluation and Policy Analysis* 6:175–87.

Levin, Henry M., and Marlaine E. Lockheed. 1991. "Creating Effective Schools." In Henry M. Levin and Marlaine E. Lockheed, eds., *Effective Schools in Developing Countries*. London: Falmer Press.

Lewis, E. Glyn, and Carolyn E. Massad. 1975. *The Teaching of English as a Foreign Language in Ten Countries. International Studies in Evaluation*, Vol. 4. Stockholm: Almquist and Wiksell.

Lincoln, Yvonna S., and Egon G. Guba. 1981. "The Place of Values in Needs Assessment." Paper presented at Annual Meeting of the American Educational Research Association, Los Angeles.

Linn, Robert L. 1983. "Testing and Instruction: Links and Distinctions." *Journal of Educational Measurement* 20:179–89.

Linn, Robert L., ed. 1989. *Educational Measurement*, 3d ed. New York: American Council on Education and Macmillan.

Lockheed, Marlaine E. 1991. "Multi-Dimensional Evaluation: Measures for both Right and Left Sides of the Equation." Paper prepared for the International Symposium of Significant Strategies to Ensure the Success of All in the Basic School, Lisbon.

_____. 1992. "World Bank Support for Capacity Building: The Challenge of Educational Assessment." PHREE Background Paper Series 92/54. World Bank, Human Development Department, Washington, D.C.

Lockheed, Marlaine E., Adriaan M. Verspoor, and associates. 1991. *Improving Primary Education in Developing Countries*. New York: Oxford University Press.

Loxley, William. 1992. "Managing International Survey Research." *Prospects* 22:289–96.

Madaus, George F., and Vincent Greaney. 1985. "The Irish Experience in Competency Testing: Implications for American Education." *American Journal of Education* 93:268–94.

Madaus, George F., and Thomas Kellaghan. 1992. "Curriculum Evaluation and Assessment." In Phillip W. Jackson, ed., *Handbook of Research on Curriculum*. New York: Macmillan.

_____. 1993. "British Experience with 'Authentic' Testing." *Phi Delta Kappan* 74:458–69.

Mauritius Examinations Syndicate. 1995. *A Survey of 9-Year-Old Children in Mauritian Schools in Literacy, Numeracy and Life Skills*. Joint UNESCO/UNICEF Project on Monitoring Education for All Goals. Reduit.

McEwen, N. 1992. "Quality Criteria for Maximizing the Use of Research." *Educational Researcher* 22 (7):20–22, 27–32.

Medrich, E. A., and J. E. Griffith. 1992. *International Mathematics and Science Assessments: What Have We Learned?* Washington, D.C.: United States Department of Education, Office of Educational Research and Improvement.

Mehrens, William A. 1992. "Using Performance Assessment for Accountability Purposes." *Educational Measurement Issues and Practice* 11 (1):3–9, 20.

Meisels, Samuel J., Aviva Dorfman, and Dorothy Steele. 1995. "Equity and Excellence in Group-Administered and Performance-Based Assessments." In Michael T. Nettles and Arie L. Nettles, eds., *Equity and Excellence in Educational Testing and Assessment*. Boston: Kluwer.

Messick, Samuel. 1984. "The Psychology of Educational Measurement." *Journal of Educational Measurement* 21:215–37.

Mullis, Ina V. S., and Lynn B. Jenkins. 1988. *The Science Report Card: Elements of Risk and Recovery, Trends and Achievement Based on the 1986 Assessment*. Princeton, N.J.: National Assessment of Educational Progress, Educational Testing Service.

_____. 1990. *The Reading Report Card, 1971 to 1988: Findings from the Nation's Report Card*. Princeton, N.J.: National Assessment of Educational Progress, Educational Testing Service.

Mullis, Ina V. S., J. A. Dossey, Eugene O. Owen, and Gary W. Phillips. 1993. *NAEP 1992: Mathematics Report Card for the Nation and the States.* Washington, D.C.: United States Department of Education, Office of Educational Research and Improvement.

Mullis, Ina V. S., J. A. Dossey, J. R. Campbell, C. A. Gentile, C. O. O' Sullivan, and A. S. Latham. 1994. *NAEP 1992: Trends in Academic Progress.* Washington, D.C.: United States Department of Education, Office of Educational Research and Improvement.

Murphy, Paud, Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas, eds. 1996. *National Assessments: Testing the System.* EDI Learning Resource Series. Washington, D.C.: World Bank.

Namibia. Ministry of Education and Culture. 1994. *How Much Do Namibia's Children Learn in School? Findings from the 1992 National Learner Baseline Assessment.* Windhoek: New Namibia Books.

Nuttall, Desmond L. 1990. "Proposals for a National System of Assessment in England and Wales." *International Journal of Educational Research* 14:373–81.

Oakes, Jennie. 1989. "What Educational Indicators? The Case for Assessing the School Context." *Educational Evaluation and Policy Analysis* 11:181–99.

Odden, Allan. 1990. "Educational Indicators in the United States: The Need for Analysis." *Educational Researcher* 19 (5):24–29.

Olmstead, P., and David Weikart. 1989. *How Nations Serve Young Children. Profile of Child Care and Education in Fourteen Countries.* Ypsilanti, Michigan: High Scope/Educational Research Foundation Press.

Organisation for Economic Co-operation and Development (OECD). 1995. *Education at a Glance. OECD Indicators.* Paris.

Owen, Eugene, G. Douglas Hodgkinson, and Albert Tuijnman. 1995. "Towards a Strategic Approach to Developing International Indicators of Student Achievement." In *Measuring What Students Learn. Mesurer les Résultats Scolaires.* Paris: Organisation for Economic Co-operation and Development.

Pelgrum, Hans, and Tjeerd Plomp. 1991. *The Use of Computers in Education Worldwide.* Oxford: Pergamon.

Phillips, Gary W. 1991. "Benefits of State-by-State Comparisons." *Educational Researcher* 20 (3):17–19.

Phillips, Gary W., Ina V. S. Mullis, M. L. Bourque, P. L. Williams, Ronald K. Hambleton, Eugene H. Owen, and P. E. Barton. 1993. *Interpreting NAEP Scales.* Washington, D.C.: National Center for Educational Statistics.

Plomp, Tjeerd. 1993. "Working Together with International Education Researchers and Policy Makers." In W. A. Hayes, ed., *Activities, Institutions and People: IEA Guidebook 1993–94.* The Hague: International Association for the Evaluation of Educational Achievement.

Postlethwaite, T. Neville, and Kenneth N. Ross. 1992. *Effective Schools in Reading. Implications for Educational Planners.* The Hague: International Association for the Evaluation of Educational Achievement.

Postlethwaite, T. Neville, and David E. Wiley. 1992. *The IEA Study of Science II: Science Achievement in Twenty-three Countries.* Oxford: Pergamon.

Prawalpruk, K. 1996. "National Assessment in Thailand." In Paud Murphy, Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas, eds., *National Assessments: Testing the System.* EDI Learning Resource Series. Washington, D.C.: World Bank.

Rojas, Carlos. 1996. "The Colombian Education Assessment System." In Paud Murphy, Vincent Greaney, Marlaine E. Lockheed, and Carlos Rojas, eds., *National Assessments: Testing the System*. Washington, D.C.: World Bank.

Ross, Kenneth N. 1987. "Sample Design." *International Journal of Educational Research* 11:57–75.

Rotberg, Iris C. 1991. "Myths in International Comparisons of Science and Mathematics Achievement." *The Bridge* 21:3–10.

Shepard, Lorrie. 1980. "Reporting the Results of State-Wide Assessment." *Studies in Educational Evaluation* 6:119–25.

Smith, M. S., J. O'Day, and D. K. Cohen. 1990. "National Curriculum American Style: Can It Be Done ? What Might It Look Like?" *American Educator* 14 (4):10–17, 40–47.

Thurow, Lester. 1992. *Head to Head. The Coming Economic Battle among Japan, Europe, and America*. New York: Morrow.

Torney, Judith V., A. N. Oppenheim, and Russell F. Farnen. 1976. *Civic Education in Ten Countries: An Empirical Study*. Stockholm: Almquist and Wiksell.

Torney-Purta, Judith. 1990. "International Comparative Research in Education: Its Role in Educational Improvement in the U.S." *Comparative Education Review* 31:32–35.

Townshend, John. 1996. "Comparing Performance Standards in Education." In Bill Boyle and Tom Christie, eds., *Issues in Setting Standards: Establishing Comparabilities*. London: Falmer.

Travers, Kenneth I., and Ian Westbury, eds. 1989. *The IEA Study of Mathematics I. Analysis of Mathematics Curricula*. Oxford: Pergamon.

UNESCO (United Nations Educational, Scientific and Cultural Organization). 1990a. *Basic Education and Literacy: World Statistical Indicators*. Paris.

_____. 1990b. *World Declaration on Education for All. Meeting Basic Learning Needs*. Adopted by the World Conference on Education for All. New York.

_____. 1994. *Monitoring Education Goals for All: Focussing on Learning Achievement. Progress Report on the Project's First Five Countries*. Paris.

U.S. Department of Education, National Center for Education Statistics. 1994. *Reading Literacy in the United States, Technical Report*. Washington, D.C.

University of Illinois. 1993. *The Standards Project for English Language Arts*. Champaign, Ill.: Center for the Study of Reading.

Visalberghi, Aldo. 1990. "Support and Venue of the Bologna Conference." *International Journal of Educational Research* 14:323–24.

Willmott, Alan S. 1977. *CSE and GCE Grading Standards: The 1973 Comparability Study*. London: Macmillan.

Windham, Douglas M. 1992. *Education for All: The Requirements*. World Conference on Education for All Monograph III. Paris: UNESCO.

World Bank. 1991. *Vocational and Technical Education and Training*. A World Bank Policy Paper. Washington, D.C.

_____. 1995a. *World Development Report: Workers in an Integrating World*. New York: Oxford University Press.

_____. 1995b. *Priorities and Strategies for Education: A World Bank Review*. Development in Practice Series. Washington, D.C.

Appendix
National Assessment Checklist

Activity	*Responsible body*	*Importance*	*Status*
1 Steering committee established	Ministry of education	Desirable	❑
2 Implementing agency appointed	Ministry of education/ steering committee	Essential	❑
3 Support of educational administration	Steering committee	Desirable	❑
4 Support of minister for education	Steering committee	Desirable	❑
5 Adequate funding	Ministry of education	Essential	❑
6 Support of teaching organizations	Steering committee	Desirable	❑
7 Terms of reference for the implementing agency	Steering committee	Desirable	❑
8 Policy informational needs addressed	Steering committee	Desirable	❑
9 Review existing relevant information	Steering committee	Desirable	❑
10 Definition of population to be assessed	Ministry of education/ steering committee	Essential	❑
11 Precise objectives identified	Implementing agency	Essential	❑
12 Assessment measures developed	Implementing agency	Essential	❑

(Appendix continues on the following page.)

Appendix *(continued)*

Activity	*Responsible body*	*Importance*	*Status*
13 Assessment measures reviewed	Implementing agency	Desirable	❑
14 Validity of measures established	Implementing agency	Desirable	❑
15 Target sample selected	Implementing agency	Essential	❑
16 Support of schools secured	Implementing agency	Essential	❑
17 Administration manual prepared	Implementing agency	Essential	❑
18 Assessment instruments printed	Implementing agency	Essential	❑
19 Instruments and manuals distributed	Implementing agency	Essential	❑
20 Supervisors/coordinators trained	Implementing agency	Essential	❑
21 Administrators trained	Implementing agency	Essential	❑
22 Data base prepared	Implementing agency	Essential	❑
23 Instruments collected and returned	Implementing agency	Essential	❑
24 Data cleaned and prepared for analysis	Implementing agency	Essential	❑
25 Analysis completed	Implementing agency	Essential	❑
26 Draft report prepared	Implementing agency	Essential	❑
27 Draft report reviewed and discussed	Implementing agency	Essential	❑
28 Final report prepared	Implementing agency	Essential	❑
29 Results disseminated	Various agencies	Desirable	❑

Directions in Development

Begun in 1994, this series contains short essays, written for a general audience, often to summarize published or forthcoming books or to highlight current development issues.

Africa's Management in the 1990s and Beyond: Reconciling Indigenous and Transplanted Institutions

Building Human Capital for Better Lives

Class Action: Improving School Performance in the Developing World through Better Health and Nutrition

Decentralization of Education: Community Financing

Decentralization of Education: Politics and Consensus

Decentralization of Education: Teacher Management

Deep Crises and Reform: What Have We Learned?

Early Child Development: Investing in the Future

Financing Health Care in Sub-Saharan Africa through User Fees and Insurance

Global Capital Supply and Demand: Is There Enough to Go Around?

Implementing Projects for the Poor: What Has Been Learned?

Improving Early Childhood Development: An Integrated Program for the Philippines

India's Family Welfare Program: Moving to a Reproductive and Child Health Approach (with a separate supplement)

Investing in People: The World Bank in Action

Managing Commodity Booms — and Busts

Meeting the Infrastructure Challenge in Latin America and the Caribbean

Monitoring the Learning Outcomes of Education Systems

MIGA: The First Five Years and Future Challenges

(continued on the following page)

Directions in Development (*continued*)

Nurturing Development: Aid and Cooperation in Today's Changing World

Nutrition in Zimbabwe: An Update

Poverty Reduction in South Asia: Promoting Participation of the Poor

Private and Public Initiatives: Working Together for Health and Education

Private Sector Participation in Water Supply and Sanitation in Latin America

Reversing the Spiral: The Population, Agriculture, and Environment Nexus in Sub-Saharan Africa (with a separate supplement)

A Strategy for Managing Water in the Middle East and North Africa

Taxing Bads by Taxing Goods: Pollution Control with Presumptive Charges

Toward Sustainable Management of Water Resources

Trade Performance and Policy in the New Independent States

The Transition from War to Peace in Sub-Saharan Africa

Unshackling the Private Sector: A Latin American Story

The Uruguay Round: Widening and Deepening the World Trading System

THE WORLD BANK
A partner in strengthening economies
and expanding markets
to improve the quality of life
for people everywhere,
especially the poorest

HEADQUARTERS

1818 H Street, N.W.
Washington, D.C. 20433, U.S.A.

TELEPHONE:	(202) 477-1234
FACSIMILE:	(202) 477-6391
TELEX:	MCI 64145 WORLDBANK
	MCI 248423 WORLDBANK
CABLE ADDRESS:	INTBAFRAD
	WASHINGTONDC
INTERNET:	http://www.worldbank.org//

EUROPEAN OFFICE

66, avenue d'Iéna
75116 Paris, France

TELEPHONE:	(1) 40.69.30.00
FACSIMILE:	(1) 40.69.30.66
TELEX:	640651

TOKYO OFFICE

Kokusai Building
1-1, Marunouchi 3-chome
Chiyoda-ku, Tokyo 100, Japan

TELEPHONE:	(3) 3214-5001
FACSIMILE:	(3) 3214-3657
TELEX:	26838

COVER DESIGN: THE MAGAZINE GROUP

ISBN 0-8213-3734-3