National Assessments of Educational Achievement

41789

VOLUME 1

Assessing National Achievement Levels in Education

Vincent Greaney
Thomas Kellaghan

THE WORLD BANK

Assessing National Achievement Levels in Education

National Assessments of Educational Achievement

VOLUME 1

Assessing National Achievement Levels in Education

Vincent Greaney
Thomas Kellaghan

THE WORLD BANK

1818 H Street NW
Washington, DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org
E-mail: feedback@worldbank.org

1 2 3 4 10 09 08 07

Cover design: Naylor Design, Washington, DC

ISBN-13: 978-0-8213-7258-6
eISBN: 978-0-8213-7259-3
DOI: 1596/978-0-8213-7258-6

Library of Congress Cataloging-in-Publication Data

Assessing national achievement levels in education / [edited by] Vincent Greaney and Thomas Kellaghan.
p. cm.
Includes bibliographical references.
ISBN 978-0-8213-7258-6 (alk. paper) — ISBN 978-0-8213-7259-3
1. Educational tests and measurements. 2. Educational evaluation. I. Greaney, Vincent. II. Kellaghan, Thomas.
LB3051.A7663 2007
371.26'2—dc22
2007022161

CONTENTS

BOXES

FIGURES

TABLES

PREFACE

In a speech to mark the first 100 days of his presidency of the World Bank Group, Robert Zoellick outlined six strategic themes to guide the Bank's work in promoting an inclusive and sustainable globalization. One of those themes focused on the role of the Bank as "a unique and special institution of knowledge and learning.... a brain trust of applied experience." Zoellick noted that this role requires the Bank "to focus continually and rigorously on results and on the assessment of effectiveness."

This challenge is greatest in education, where the large body of empirical evidence linking education to economic growth indicates that improved enrollment and completion rates are necessary, but not sufficient, conditions for poverty reduction. Instead, enhanced learning outcomes—in the form of increased student knowledge and cognitive skills—are key to alleviating poverty and improving economic competitiveness (and will be crucial for sustaining the gains achieved in education access to date). In other words, the full potency of education in relation to economic growth can only be realized if the education on offer is of high quality and student knowledge and cognitive skills are developed.

The available evidence indicates that the quality of learning outcomes in developing countries is very poor. At the same time, few of these countries systematically monitor such outcomes either through

conducting their own assessments of student achievement or through participating in regional or international assessments. The lack of this type of regular, system-level information on student learning makes it difficult to gauge overall levels of achievement, to assess the relative performance of particular subgroups, and to monitor changes in performance over time. It also makes it difficult to determine the effectiveness of government policies designed to improve outcomes in these and other areas.

This is a core issue for the Bank and its client countries as the focus shifts from access to achievement. It also is an area in which there is a dearth of tools and resources suited to the needs of developing countries. This series of books, edited by Vincent Greaney and Thomas Kellaghan, contributes in a significant way to closing this gap. The series is designed to address many of the issues involved in making learning outcomes a more central part of the educational agenda in lower-income countries. It will help countries to develop capacity to measure national levels of student learning in more valid, sustainable, and systematic ways. Such capacity will hopefully translate into evidence-based policymaking that leads to observable improvement in the quality of student learning. It is an important building block toward achieving the real promise of education for dynamic economies.

Marguerite Clarke
Senior Education Specialist
The World Bank

ACKNOWLEDGMENTS

A team led by Vincent Greaney (consultant, Human Development Network, Education Group, World Bank) and Thomas Kellaghan (Educational Research Centre, St. Patrick's College, Dublin) prepared this series of books.

Other contributors to the series were Sylvia Acana (Uganda National Examinations Board), Prue Anderson (Australian Council for Educational Research), Fernando Cartwright (Canadian Council on Learning), Jean Dumais (Statistics Canada), Chris Freeman (Australian Council for Educational Research), Hew Gough (Statistics Canada), Sara Howie (University of Pretoria), George Morgan (Australian Council for Educational Research), T. Scott Murray (DataAngel Policy Research) and Gerry Shiel (Educational Research Centre, St. Patrick's College, Dublin).

The work was carried out under the general direction of Ruth Kagia, World Bank Education Sector Director, and Robin Horn, Education Sector Manager. Robert Prouty initiated and supervised the project up to August 2007. Marguerite Clarke supervised the project in its later stages through review and publication. We are grateful for contributions of the review panel: Al Beaton (Boston College), Irwin Kirsch (Educational Testing Service), and Benoit Millot (World Bank).

Additional peer-review comments were provided by a number of World Bank staff, including Carlos Rojas, Eduardo Velez, Elizabeth King, Harry Patrinos, Helen Abadzi, Jee-Peng Tan, Marguerite Clarke, Maureen Lewis, Raisa Venalainen, Regina Bendokat, Robert Prouty, and Robin Horn.

Special thanks are due to Aidan Mulkeen and to Sarah Plouffe. We received valuable support from Cynthia Guttman, Matseko Ramokoena, Aleksandra Sawicka, Pam Spagnoli, Beata Thorstensen, Myriam Waiser, Peter Winograd, and Hans Wagemaker. We are also grateful to Patricia Arregui, Harsha Aturupane, Luis Benveniste, Jean-Marc Bernard, Carly Cheevers, Zewdu Gebrekidan, Venita Kaul, Pedro Ravela, and Kin Bing Wu.

We wish to thank the following institutions for permission to reproduce material: Examinations Council of Lesotho, International Association for the Evaluation of Educational Achievement, National Center for Education Statistics of the U.S. Department of Education, the Organisation for Economic Co-operation and Development, and the Papua New Guinea Department of Education.

Hilary Walshe helped prepare the manuscript. Book design, editing, and production were coordinated by Mary Fisk and Paola Scalabrin of the World Bank's Office of the Publisher.

The Irish Educational Trust Fund; the Bank Netherlands Partnership Program; the Educational Research Centre, Dublin; and the Australian Council for Educational Research have generously supported preparation and publication of this series.

ABBREVIATIONS

CONFEMEN	Conférence des Ministres de l'Education des Pays ayant le Français en Partage
DiNIECE	Dirección Nacional de Información y Evaluación de la Calidad Educativa (Argentina)
EFA	Education for All
IEA	International Association for the Evaluation of Educational Achievement
IIEP	International Institute for Educational Planning
LLECE	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación
MOE	ministry of education
MESyFOD	Modernización de la Educación Secundaria y Formación Docente (Uruguay)
NAEP	National Assessment of Educational Progress (United States)
NAPE	National Assessment of Progress in Education (Uganda)
NSC	national steering committee
OECD	Organisation for Economic Co-operation and Development
PASEC	Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN
PIRLS	Progress in International Reading Literacy Study

PISA	Programme for International Student Assessment
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SIMCE	Sistema de Medición de la Calidad de la Educación (Chile)
SNED	National System of Teacher Performance Assessment in Publicly Supported Schools (Chile)
SSA	Sarva Shiksha Abhiyan (India)
TA	technical assistance
TIMSS	Trends in International Mathematics and Science Study
UMRE	Unidad de Medición de Resultados Educativos (Uruguay)
UNEB	Uganda National Examinations Board
UNESCO	United Nations Educational, Scientific, and Cultural Organization

CHAPTER 1

INTRODUCTION

In this introductory book, we describe the main features of national and international assessments, both of which became extremely popular tools for determining the quality of education in the 1990s and 2000s. This increase in popularity reflects two important developments. First, it reflects increasing globalization and interest in global mandates, including Education for All (UNESCO 2000). Second, it represents an overall shift in emphasis in assessing the quality of education from a concern with inputs (such as student participation rates, physical facilities, curriculum materials, and teacher training) to a concern with outcomes (such as the knowledge and skills that students have acquired as a result of their exposure to schooling) (Kellaghan and Greaney 2001b). This emphasis on outcomes can, in turn, be considered an expression of concern with the development of human capital in the belief (a) that knowledge is replacing raw materials and labor as resources in economic development and (b) that the availability of human knowledge and skills is critical in determining a country's rate of economic development and its competitiveness in an international market (Kellaghan and Greaney 2001a). A response to this concern has required information on the performance of education systems, which, in turn, has involved a shift from the traditional use of

achievement tests to assess individual students toward their use to obtain information about the achievements of the system of education as a whole (or a clearly defined part of the system).

The development of national assessment capacity has enabled ministries of education—as part of their management function—to describe national levels of learning achievement, especially in key subject areas, and to compare achievement levels of key subgroups (such as boys and girls, ethnic groups, urban and rural students, and public and private school students). It has also provided evidence that enables ministries to support or refute claims that standards of student achievement are rising or falling over time.

Despite growth in national and international assessment activity, a lack of appreciation still exists in many quarters about the potential value of the data that assessments can provide, as well as a deficit in the skills required to carry out a technically sound assessment. Even when countries conduct a national assessment or participate in an international one, the information yielded by the assessment is frequently not fully exploited. A number of reasons may account for this: the policy makers may have been only peripherally involved in the assessment and may not have been fully committed to it; the results of analyses may not have been communicated in a form that was intelligible to policy makers; or the policy makers may not have fully appreciated the implications of findings for social policy in general or for educational policy in particular relating to curricular provision, the allocation of resources, the practice of teaching, and teachers' professional development.

This series of books is designed to address such issues by introducing readers to the complex technology that has grown up around the administration of national and international assessments. This introductory book describes key national assessment concepts and procedures. It is intended primarily for policy makers and decision makers in education. The purposes and main features of *national assessments* are described in chapter 2 (see also appendix A). The reasons for carrying out a national assessment are considered in chapter 3, and the main decisions that have to be made in the design and planning of an assessment are covered in chapter 4. Issues (as well as common errors) to be borne in mind in the design, implementation, analysis,

reporting, and use of a national assessment are identified in chapter 5. In chapter 6, *international assessments* of student achievement, which share many procedural features with national assessments (such as sampling, administration, background data collected, and methods of analysis—see appendix B), are described.

The main point of difference between national and international assessments highlights both a strength and a weakness of an international assessment. The strength is that an international assessment provides data from a number of countries, thereby allowing each country to compare the results of its students with the results achieved by students in other countries. The weakness is that the requirement that test instruments be acceptable in all participating countries means that they may not accurately reflect the range of achievements of students in individual countries.

A further feature of international assessments is that many participating countries carry out internal analyses that are based on data collected within a country. Thus, the data collected for the international study can be used for what is, in effect, a national assessment. However, the practice is not without its problems, and the data that are collected in this way may be less appropriate for policy than if they had been collected for a dedicated national assessment.

An intermediate procedure that lies between national assessments in individual countries and large-scale international studies that span the globe is the *regional study* in which a number of countries in a region that may share many socioeconomic and cultural features collaborate in a study (see appendix C).

A further variation is a *subnational assessment* in which an assessment is confined to a region (a province or state) within a country. Subnational assessments have been carried out in a number of large countries (such as Argentina, Brazil, and the United States) to meet local or regional information needs. Those exercises are relatively independent and differ from national assessments in that participants in all regions within a country do not respond to the same instruments and procedures; thus, direct comparisons of student achievement between regions are not possible.

In the final chapter of this volume, some overall conclusions are presented, together with consideration of conditions relating to the

development and institutionalization of national assessment capacity and to the optimal use of assessment findings. At the end of the book, the main features of national assessments in nine countries are described (appendix A), followed by descriptions of three international studies (appendix B) and three regional studies (appendix C).

Subsequent books in this series provide details of the design and implementation of a national assessment. The books are designed to provide those directly involved in the tasks of constructing tests and questionnaires and of collecting, analyzing, or describing data in a national assessment with an introduction to—and basic skills in—key technical aspects of the tasks involved.

The second book, *Developing Tests and Questionnaires for a National Assessment of Educational Achievement*, has sections on developing (a) achievement tests, (b) questionnaires, and (c) administration manuals. The first section addresses the design of achievement tests and the role that a test framework and blueprint or table of specifications plays in the design. It describes the process of item writing and gives examples of various item types, including multiple-choice, short-answer, and open-ended response items. It also describes the item review or paneling process, an essential exercise to ensure test-content validity. It includes guidelines for conducting pretests, selecting items for the final test, and producing the final version of a test. The section concludes with a brief treatment of training scorers or raters and hand-scoring test items. The second section describes steps in the construction of questionnaires: designing a questionnaire, writing items, scoring and coding responses, and linking data derived from the questionnaire with students' achievement scores. The final section describes the design and content of an administration manual and the selection and role of a test administrator. The book has an accompanying CD, which contains test and questionnaire items released from national and international assessments and a test administration manual.

Implementing a National Assessment of Educational Achievement, the third book in the series, is also divided into three sections. The first section focuses on practical issues to be addressed in implementing a large-scale national assessment program. It covers planning, budgeting, staffing, arranging facilities and equipment, contacting schools, selecting test administrators, packing and shipping, and

ensuring test security. This section also covers the logistical aspects of test scoring, data cleaning, and report writing. The second section includes a step-by-step guide designed to enable assessment teams to draw an appropriate national sample. It includes a CD with sampling software and a training dataset to be used in conjunction with the guide. Topics addressed are defining the population to be assessed, creating a sampling frame, calculating an appropriate sample size, sampling with probability proportional to size, and conducting multistage sampling. Data cleaning and data management are treated in the final section. This section is also supported by a CD with step-by-step exercises to help users prepare national assessment data for analysis. Procedures for data verification and data validation, including "wild codes" and within-file and between-file consistency checks, are described.

Analyzing Data from a National Assessment of Educational Achievement, the fourth book, is supported by two CDs, which require users to apply statistical procedures to datasets and to check their mastery levels against solutions depicted on screenshots in the text. The first half of the book deals with the generation of item-level data using both classical test and item response theory approaches. Topics addressed include analyzing pilot and final test items, monitoring change in performance over time, building a test from previously created items, equating, and developing performance or proficiency levels. The second half of the book is designed to help analysts carry out basic-level analysis of national assessment results and includes sections on measures of central tendency and dispersion, mean score differences, identification of high and low achievers, correlation, regression, and visual representation of data.

Reporting and Using Results from a National Assessment of Educational Achievement, the final book in the series, focuses on writing reports in a way that will influence policy. It introduces a methodology for designing a dissemination and communication strategy for a national assessment program. It also describes the preparation of a technical report, press releases, briefings for key policy makers, and reports for teachers and other specialist groups. The second section of the book highlights ways that countries have actually used the results of national assessments for policy making, curriculum reform, resource

allocation, teacher training, accountability, and monitoring of changes in achievement and other variables over time.

Those who study the content of these books and who carry out the specified exercises should acquire the basic skills required for a national assessment. They should, however, bear in mind three factors. First, they should not regard the books as providing simple formulas or algorithms to be applied mechanically but should be prepared to exercise judgment at varying points in the national assessment (for example, in selection of test content, in sampling, and in analysis). Judgment in these matters should improve with experience. Second, users may, on occasion, require the advice of more experienced practitioners in making their judgments. Third, users should be prepared to adapt to the changes in knowledge and technology that will inevitably occur in the coming years.

CHAPTER 2

NATIONAL ASSESSMENTS OF STUDENT ACHIEVEMENT

We begin the chapter by defining a national assessment and listing questions that a national assessment would be designed to answer. A listing of the main elements of a national assessment follows. Finally, we consider the differences between a national assessment and public examinations.

A national assessment is designed to describe the achievement of students in a curriculum area aggregated to provide an estimate of the achievement level in the education system as a whole at a particular age or grade level. It provides data for a type of national education audit carried out to inform policy makers about key aspects of the system. Normally, it involves administration of achievement tests either to a sample or to a population of students, usually focusing on a particular sector in the system (such as fifth grade or 13-year-old students). Teachers and others (for example, parents, principals, and students) may be asked to provide background information, usually in questionnaires, which, when related to student achievement, can provide insights about how achievement is related to factors such as household characteristics, levels of teacher training, teachers' attitudes toward curriculum areas, teacher knowledge, and availability of teaching and learning materials.

National assessment systems in various parts of the world tend to have common features. All include an assessment of students' language or literacy and of students' mathematics abilities or numeracy. Some systems assess students' achievements in a second language, science, art, music, or social studies. In practically all national assessment systems, students at the primary-school level are assessed. In many systems, national assessments are also carried out in secondary school, usually during the period of compulsory education.

Differences also exist in national assessment systems from country to country. First, they differ in the frequency with which assessments are carried out. In some countries, an assessment is carried out every year, although the curriculum area that is assessed may vary from year to year. In other systems, assessments are less frequent. Second, they differ in the agency that carries out an assessment. In some systems, the ministry of education carries out the assessment; in others, the assessment is by a national research center, a consortium of educational bodies, a university, or an examination board. Third, participation by a school may be voluntary or may be mandated. When voluntary, nonparticipation of some schools will almost invariably bias the results and lead to an inaccurate reflection of achievement levels in the education system.

Although most industrial countries have had systems of national assessment for some time, it was not until the 1990s that the capacity to administer assessments became more widely available in other parts of the world. For example, rapid development in the establishment of national assessments took place during the 1990s in Latin American and Caribbean counties, often to provide baseline data for educational reforms (Rojas and Esquivel 1998). The development represented a shift in the assessment of quality from emphasis on educational inputs to outcomes following the Jomtien Declaration (see *World Declaration on Education for All* 1990). Article 4 of the Jomtien Declaration states that the focus of basic education should be "on actual learning acquisition and outcome, rather than exclusively upon enrolment, continued participation in organized programs and completion of certification requirements" (*World Declaration on Education for All* 1990, 5). More recently, the Dakar Framework for Action (UNESCO 2000), which was produced at the end of the 10-year follow-up to Jomtien, again highlighted the importance of learning outcomes. Among its list of

seven agreed goals was, by 2015, to improve "all aspects of the quality of education … so that recognised and measurable outcomes are achieved by all, especially in literacy, numeracy, and essential life skills" (UNESCO 2000, iv, 7).

These statements imply that, for countries pledged to achieving the goals of Education for All (EFA), efforts to enhance the quality of education will have to be accompanied by procedures that will provide information on students' learning. As a result, national governments and donor agencies have greatly increased support for monitoring student achievement through national assessments. The assumption is frequently made not only that national assessments will provide information on the state of education, but also that use of the information should lead to improvement in student achievements. Whether this improvement ultimately happens remains to be seen. So far, the expectation that EFA and regular monitoring of achievement levels would result in an improvement in learning standards does not seem to have materialized (Postlethwaite 2004). This outcome may be because—although EFA led to rapid increases in numbers attending school—larger numbers were not matched by increased resources (especially trained teachers). Furthermore, the information obtained from assessments has often been of poor quality, and even when it has not, it has not been systematically factored into decision making.

All national assessments seek answers to one or more of the following questions:

- How well are students learning in the education system (with reference to general expectations, aims of the curriculum, preparation for further learning, or preparation for life)?
- Does evidence indicate particular strengths and weaknesses in students' knowledge and skills?
- Do particular subgroups in the population perform poorly? Do disparities exist, for example, between the achievements of (a) boys and girls, (b) students in urban and rural locations, (c) students from different language or ethnic groups, or (d) students in different regions of the country?
- What factors are associated with student achievement? To what extent does achievement vary with characteristics of the learning

environment (for example, school resources, teacher preparation and competence, and type of school) or with students' home and community circumstances?
- Are government standards being met in the provision of resources (for example, textbooks, teacher qualifications, and other quality inputs)?
- Do the achievements of students change over time? This question may be of particular interest if reforms of the education system are being undertaken. Answering the question requires carrying out assessments that yield comparable data at different points in time (Kellaghan and Greaney 2001b, 2004).

Most of those questions were addressed in the design and implementation of Ethiopia's national assessment (see box 2.1).

A feature of Vietnam's approach to national assessment, in addition to assessing student achievement, was a strong focus on key inputs, such as physical conditions in schools, access to educational materials, and teacher qualifications (see box 2.2).

BOX 2.1

Ethiopia: National Assessment Objectives

1. To determine the level of student academic achievement and attitude development in Ethiopian primary education.
2. To analyze variations in student achievement by region, gender, location, and language of instruction.
3. To explore factors that influence student achievement in primary education.
4. To monitor the improvement of student learning achievement from the first baseline study in 1999/2000.
5. To build the capacity of the education system in national assessment.
6. To create reliable baseline data for the future.
7. To generate recommendations for policy making to improve educational quality.

Source: Ethiopia, National Organisation for Examinations 2005.

BOX 2.2

Example of Questions Addressed by Vietnam's National Assessment

Questions Related to Inputs

- What are the characteristics of grade 5 pupils?
- What are the teaching conditions in grade 5 classrooms and in primary schools?
- What is the general condition of the school buildings?

Questions Related to Standards of Educational Provision

- Were ministry standards met regarding
  - Class size?
  - Classroom furniture?
  - Qualifications of staff members?

Questions Related to Equity of School Inputs

- Was there equity of resources among provinces and among schools within provinces in terms of
  - Material resource inputs?
  - Human resource inputs?

Questions Related to Achievement

- What percentage of pupils reached the different levels of skills in reading and mathematics?
- What was the level of grade 5 teachers in reading and mathematics?

Questions Related to Influences on Achievement

- What were the major factors accounting for the variance in reading and mathematics achievement?
- What were the major variables that differentiated between the most and least effective schools?

Source: World Bank 2004.

WHAT ARE THE MAIN ELEMENTS IN A NATIONAL ASSESSMENT?

Although national assessments can vary in how they are implemented, they tend to have a number of common elements (see box 2.3 and Kellaghan and Greaney 2001b, 2004).

BOX 2.3

Main Elements of a National Assessment

- The ministry of education (MOE) appoints either an implementing agency within the ministry or an independent external body (for example, a university department or a research organization), and it provides funding.
- The MOE determines policy needs to be addressed in the assessment, sometimes in consultation with key education stakeholders (for example, teachers' representatives, curriculum specialists, business people, and parents).
- The MOE, or a steering committee nominated by it, identifies the population to be assessed (for example, fourth grade students).
- The MOE determines the area of achievement to be assessed (for example, literacy or numeracy).
- The implementing agency defines the area of achievement and describes it in terms of content and cognitive skills.
- The implementing agency prepares achievement tests and supporting questionnaires and administration manuals, and it takes steps to ensure their validity.
- The tests and supporting documents are pilot-tested by the implementing agency and subsequently are reviewed by the steering committee and other competent bodies to (a) determine curriculum appropriateness and (b) ensure that items reflect gender, ethnic, and cultural sensitivities.
- The implementing agency selects the targeted sample (or population) of schools or students, arranges for printing of materials, and establishes communication with selected schools.
- The implementing agency trains test administrators (for example, classroom teachers, school inspectors, or graduate university students).
- The survey instruments (tests and questionnaires) are administered in schools on a specified date under the overall direction of the implementing agency.
- The implementing agency takes responsibility for collecting survey instruments, for scoring, and for cleaning and preparing data for analysis.

(*continued*)

BOX 2.3

- The implementing agency establishes the reliability of the assessment instruments and procedures.
- The implementing agency carries out the data analysis.
- The draft reports are prepared by the implementing agency and reviewed by the steering committee.
- The final reports are prepared by the implementing agency and are disseminated by the appropriate authority.
- The MOE and other relevant stakeholders review the results in light of the policy needs that they are meant to address and determine an appropriate course of action.

Source: Authors.

It is clear from the list of elements in box 2.3 that a good deal of thought and preparation are required before students respond to assessment tasks. A body with responsibility for collecting data must be appointed, decisions must be made about the policy issues to be addressed, and tests and questionnaires must be designed and tried out. In preparation for the actual testing, samples (or populations) of schools and of students must be identified, schools must be contacted, and test administrators must be selected and trained. In some countries (for example, India, Vietnam, and some African countries), teachers have been assessed on the tasks taken by their students (see A.1 and A.2 in appendix A and C.1 in appendix C). Following test administration, a lot of time and effort will be required to prepare data for analysis, to carry out analyses, and to write reports.

Low-income countries have to deal with problems over and above those encountered by other countries in attempting to carry out a national assessment. Education budgets may be meager. According to 2005 data (World Bank 2007), some countries devote 2 percent or less of gross domestic product to public education (for example, Bangladesh, Cameroon, Chad, the Dominican Republic, Guinea, Kazakhstan, the Lao People's Democratic Republic, Mauritania, Pakistan, Peru, the Republic of Congo, United Arab Emirates, and Zambia) compared to more than 5 percent in most middle- and high-income countries.

Competing demands within the education sector for activities such as school construction, teacher training, and provision of educational materials can result in nonavailability of funds for monitoring educational achievement. Furthermore, many low- and, indeed, middle-income countries have weak institutional capacity for carrying out a national assessment. They may also have to face additional administrative and communication problems caused by inadequate roads, mail service, and telephone service. Finally, the very high between-school variation in student achievement found in some low-income countries requires a large sample (see UNEB 2006; World Bank 2004).

HOW DOES A NATIONAL ASSESSMENT DIFFER FROM PUBLIC EXAMINATIONS?

Public examinations play a crucial role in many education systems in certifying student achievement, in selecting students for further study, and in standardizing what is taught and learned in schools. Sometimes, public examinations are thought to provide the same information as a national assessment, thus appearing to eliminate the need for a national assessment system in a country that has a public examination system. However, public examinations cannot provide the kind of information that a national assessment seeks to provide.

First, since public examinations play a major role in selecting students (for the next highest level in the education system and sometimes for jobs), they seek to discriminate between relatively high achieving students and so may not provide adequate coverage of the curriculum. Second, examinations, as well as the characteristics of students who take them, change from year to year, thereby limiting the inferences that can be made from comparisons over time. Third, the fact that "high stakes" are attached to performance (that is, how students do on an examination has important consequences for them and perhaps for their teachers) means that teachers (and students) may focus on those areas of the curriculum that are examined to the neglect of important areas that are not examined (for example, practical skills), so that performance on the examination does not provide

TABLE 2.1

Differences between National Assessments and Public Examinations

	National assessments	Public examinations
Purpose	To provide feedback to policy makers.	To certify and select students.
Frequency	For individual subjects offered on a regular basis (such as every four years).	Annually and more often where the system allows for repeats.
Duration	One or two days.	Can extend over a few weeks.
Who is tested?	Usually a sample of students at a particular grade or age level.	All students who wish to take this examination at the examination grade level.
Format	Usually multiple choice and short answer.	Usually essay and multiple choice.
Stakes: importance for students, teachers, and others	Low importance.	Great importance.
Coverage of curriculum	Generally confined to one or two subjects.	Covers main subject areas.
Effect on teaching	Very little direct effect.	Major effect: teacher tendency to teach what is expected on the examination.
Additional tuition sought for students	Very unlikely.	Frequently.
Do students get results?	Seldom.	Yes.
Is additional information collected from students?	Frequently, in student questionnaires.	Seldom.
Scoring	Usually involves statistically sophisticated techniques.	Usually a simple process that is based on a predetermined marking scheme.
Effect on level of student attainment	Unlikely to have an effect.	Poor results or the prospect of failure, which can lead to early dropout.
Usefulness for monitoring trends in achievement levels over time	Appropriate if tests are designed with monitoring in mind.	Not appropriate because examination questions and candidate populations change from year to year.

Source: Authors.

an accurate reflection of the intended curriculum. Although there are some exceptions, decisions about individual students, teachers, or schools are not normally made following a national assessment.

Fourth, information on student achievement is usually required at an earlier age than that at which public examinations are held. Fifth, the kind of contextual information (about teaching, resources, and students and their homes) that is used in the interpretation of achievement data collected in national assessments is not available to interpret public examination results (Kellaghan 2006). Table 2.1 summarizes the major differences between national assessments and public examinations.

CHAPTER 3

WHY CARRY OUT A NATIONAL ASSESSMENT?

A decision to carry out a national assessment might be made for a variety of reasons. Frequently, national assessments reflect the efforts of a government to "modernize" its education system by introducing a business management (corporatist) approach (Kellaghan 2003). This approach draws on concepts used in the world of business, such as strategic planning and a focus on deliverables and results, and it may involve accountability based on performance. Viewed from this perspective, a national assessment is a tool for providing feedback on a limited number of outcome measures that are considered important by policy makers, politicians, and the broader educational community.

A key objective of this approach is to provide information on the operation of the education system. Many governments lack basic information on aspects of the system—especially student achievement levels—and even on basic inputs to the system. National assessments can provide such information, which is a key prerequisite for sound policy making. For example, Vietnam's national assessment helped establish that many classrooms lacked basic resources (World Bank 2004). In a similar vein, Zanzibar's assessment reported that 45 percent of pupils lacked a place to sit (Nassor and Mohammed 1998). Bhutan's national assessment noted that some students had to

spend several hours each day traveling to and from school (Bhutan, Board of Examinations, Ministry of Education 2004). Namibia's assessment showed that many teachers had limited mastery of basic skills in English and mathematics (Makuwa 2005).

The need to obtain information on what students learn at school has assumed increasing importance with the development of the so-called knowledge economy. Some analysts argue that students will need higher levels of knowledge and skills—particularly in the areas of mathematics and science—than in the past if they are to participate meaningfully in the world of work in the future. Furthermore, because ready access to goods and services increases with globalization, a country's ability to compete successfully is considered to depend to a considerable degree on the skills of workers and management in their use of capital and technology. This factor might point to the need to compare the performance of students in one's education system with the performance of students in other systems, although a danger exists in assigning too much importance to aggregate student achievement in accounting for economic growth, given the many other factors involved (Kellaghan and Greaney 2001a).

National assessments, when administered over a period of time, can be used to determine whether standards improve, deteriorate, or remain static. Many developing countries face the problem of expanding enrollments, building many new schools, and training large numbers of teachers while at the same time trying to improve the quality of education—sometimes against a background of a decreased budget. In this situation, governments need to monitor achievement levels to determine how changes in enrollment and budgetary conditions affect the quality of learning. Otherwise, the risk exists that increased enrollment rates may be readily accepted as evidence of an improvement in the quality of education.

National assessment data have been used to monitor achievement over time. A series of studies in Africa between 1995/96 and 2000/01 revealed a significant decline in reading literacy scores in Malawi, Namibia, and Zambia (see figure C.1.2 in appendix C). In the United States, the National Assessment of Educational Progress (NAEP), which has monitored levels of reading achievement over almost three decades, found that although nine-year-old black and Hispanic

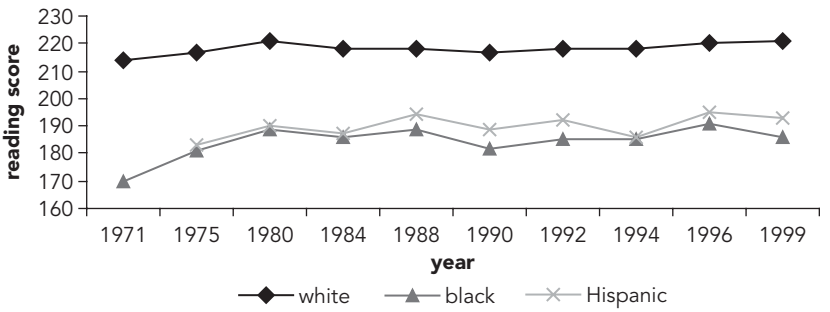
children reduced the achievement gap with whites up to about 1980, the test score differential remained fairly constant thereafter (figure 3.1). Also in the United States, the NAEP helped identify the changing levels of reading achievement in various states (figure 3.2). In Nepal, results of national assessments were used to monitor (a) changes in achievement over the period 1997–2001 and, in particular, (b) effects of policy decisions relating to budget, curricula, textbooks, teaching materials, and teacher development (see A.6 in appendix A).

When national assessment data are used to monitor achievement over time, the same test should be used in each assessment or, if different tests are used, some items should be common, so that performance on the tests can be equated or linked. In either case, the common items should be kept secure so that student or teacher familiarity with their content does not invalidate the comparisons being made.

Other uses that can be made of a national assessment depend on whether data were collected in a sample of schools or in a census in which information is obtained about all (or most) schools. In both cases, results can be used to provide direction to policy makers who are interested in enhancing educational quality. For example, the results can help governments identify the strength of the association between the quality of student learning and various factors over which they have some control (for example, availability of textbooks, class size, and number of years of teacher preservice training).

FIGURE 3.1

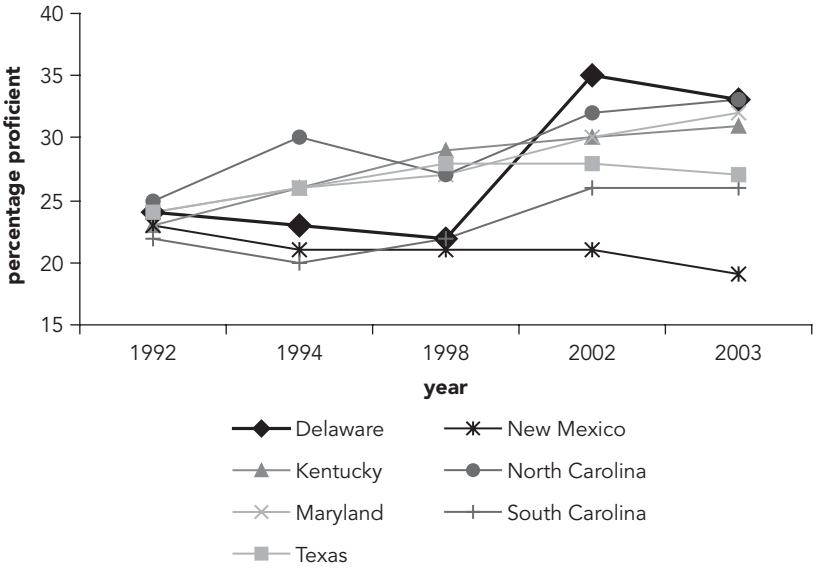
The Achievement Gap in the United States for Nine-Year-Old Students: NAEP Reading Assessment, 1971–99



Source: Winograd and Thorstensen 2004.

FIGURE 3.2

Percentages of Fourth Grade Students at or above "Proficient" in Reading, NAEP 1992–2003



Source: Winograd and Thorstensen 2004.

An analysis of findings can lead to decisions affecting the provision of resources in the education system in general (for example, for the reform of curricula and textbooks or for teacher development) or in categories of schools with particular characteristics (for example, schools in rural areas or schools serving students in socioeconomically disadvantaged areas). Many examples can be found of the use of the findings of national and international assessments for such purposes. They have been used in Australia to provide programs designed to increase the participation and performance of girls in mathematics and science (Keeves 1995); they have prompted curriculum reform in low- and middle-income countries (Elley 2005), have helped divert financial resources to poorer schools in Chile (see A.7 in appendix A), and have promoted teacher professionalism in Uruguay (see A.3 in appendix A).

The results of a national assessment may also be used to change practice in the classroom (Horn, Wolff, and Velez 1992). Getting

information to teachers and effecting changes in their behavior that will substantially raise the achievements of students, however, is not an easy task. The pressure on schools and classrooms to change is greatest when the results of a national assessment are based on a census, not a sample, and when high stakes are attached to performance. No specific action may be taken by the authorities apart from the publication of information about performance (for example, in league tables), or sanctions may be attached to performance. Sanctions can take the form of rewards for improved performance (for example, schools, teachers, or both receive economic incentives if students achieve a specific target) or "punishment" for poor performance (for example, nonpromotion of students or dismissal of teachers) (see A.7 in appendix A for a brief description of Chile's reward program).

When a national assessment obtains information about the achievement of students in all (or most) schools, some policy makers may see an opportunity to use these data to judge the quality of teachers and schools. Obviously, teachers and students should bear some responsibility for learning, but the role of institutions, agencies, and individuals that exercise control over the resources and activities of schools should also be reflected in an accountability system. Apportioning fairly the responsibilities of all stakeholders is important, whether an assessment is sample-based or census-based. The national assessment in Uruguay provides a good example of recognition of the responsibility of a variety of stakeholders (including the state) for student achievement (see A.3 in appendix A).

In some cases, a national assessment may simply have a symbolic role, which is designed to legitimate state action by embracing internationally accepted models of modernity and by imbuing the policy-making process with the guise of scientific rationality (Benveniste 2000, 2002; Kellaghan 2003). When this role motivates a national assessment, the act of assessment has greater significance than its outcomes. If a national assessment is carried out simply to meet the requirement of a donor agency, or even to meet a government's international commitments to monitor progress toward achieving the Millennium Development Goals, it may have little more than symbolic value, and its findings may not be seriously considered in the management of the education system or in policy making.

CHAPTER 4

DECISIONS IN A NATIONAL ASSESSMENT

In this chapter, we consider 12 decisions that are involved in planning a national assessment (see Greaney and Kellaghan 1996; Kellaghan 1997; and Kellaghan and Greaney 2001b, 2004).

WHO SHOULD GIVE POLICY GUIDANCE FOR THE NATIONAL ASSESSMENT?

The ministry of education should appoint a national steering committee (NSC) to provide overall guidance to the agency that will carry out the assessment. The committee can help ensure that the national assessment has status and that key policy questions of interest to the ministry and others are addressed. It could also help resolve serious administrative and financial problems that might arise during the implementation of the national assessment. Giving the NSC a degree of ownership over the direction and intent of the national assessment also increases the likelihood that the results of the assessment will play a role in future policy making.

The composition of an NSC will vary from country to country, depending on the power structure within the education system. In

addition to representatives of the ministry of education, NSCs might include representatives of major ethnic, religious, and linguistic groups, as well as those groups whose members will be expected to act on the results (such as teacher trainers, teachers, school inspectors, and curriculum personnel). Box 4.1 lists suggested members of a steering committee for a national assessment in Sierra Leone proposed by participants at an international workshop. Addressing the information needs of those various stakeholders should help ensure that the national assessment exercise does not result in a report that is criticized or ignored because of its failure to address the "correct" questions.

The NSC should not be overburdened with meetings and should not be required to address routine implementation tasks related to the national assessment. In some cases, the NSC may provide direction at the initial stage by identifying the purpose of and rationale for the assessment, by determining the curriculum areas and grade levels to be assessed, or by selecting the agency or agencies to conduct the assessment, although those items may also be decided before the committee is established. The NSC is likely to be most active at the

BOX 4.1

Proposed NSC Membership in Sierra Leone

- Basic Education Commission
- Civil Society Movement
- Decentralized Secretariat
- Director-General of Education (chair)
- Education Planning Directorate
- Inter-Religious Council
- National Curriculum Research Development Centre
- Sierra Leone Teachers Union
- Statistics Sierra Leone
- Teacher Training Colleges
- West African Examinations Council

start of the assessment exercise, whereas the implementing agency will be responsible for most of the detailed work, such as instrument development, sampling, analysis, and reporting. The implementing agency, however, should provide the NSC with draft copies of tests and questionnaires and with descriptions of proposed procedures so that committee members can provide guidance and can ensure that the information needs that prompted the assessment in the first place are being adequately addressed. NSC members should also review draft reports prepared by the implementing agency.

Responsibility for providing policy guidance: Ministry of education

WHO SHOULD CARRY OUT THE NATIONAL ASSESSMENT?

A national assessment should be carried out by a credible team or organization whose work can command respect and enhance the likelihood of broad-scale acceptance of the findings. Various countries have assigned responsibility for national assessments to groups ranging from teams set up within the ministry of education, to autonomous bodies (universities, research centers), to nonnational technical teams. We would expect a variety of factors to influence such a decision, including levels of national technical capacity, as well as administrative and political circumstances. Table 4.1 lists some potential advantages and disadvantages of different categories of implementation agencies that merit consideration in deciding who should carry out an assessment.

In some cases, traditions and legislation may impose restrictions on the freedom of a ministry of education in choosing an implementing agency. In Argentina, for example, provinces must authorize the curricular contents to be evaluated in the national assessment. Initially, provinces were asked to produce test items; however, many provinces lacked the technical capacity to do so. At a later stage, provinces were presented with a set of sample questions for their endorsement and the Dirección Nacional de Información y Evaluación de la Calidad Educativa (DiNIECE) constructed the final assessment instruments from the pool of preapproved test items. More recently, test items have been designed independently by university personnel and approved by the national Federal Council. The DiNIECE remains

TABLE 4.1

Options for Implementing a National Assessment

Designated agency	Advantages	Disadvantages
Drawn from staff of ministry of education	Likely to be trusted by ministry. Enjoys ready access to key personnel, materials, and data (for example, school population data). Funds that may not have to be secured for staff time.	Findings might be subject to political manipulation including suppression. May be viewed skeptically by other stakeholders. Staff who may be required to undertake many other tasks. Technical capacity who may be lacking.
Drawn from staff of public examination unit	Usually is credible. Has experience in running secure assessments. Funds that may not have to be secured for staff time. Some skills (for example, test development) that can be transferred to enhance the examination unit. More likely to be sustainable than some other models.	Staff who may be required to undertake many other tasks. Technical capacity that may be weak. May lack ready access to data. Public examination experience that may result in test items that are too difficult.
Drawn from research/ university sector	Findings that may be more credible with stakeholders. Greater likelihood of some technical competence. May use data for further studies of the education system.	Have to raise funds to cover staff costs. May be less sustainable than some other models. May come into conflict with education ministry.

Designated agency	Advantages	Disadvantages
Recruited as foreign technical assistance (TA)	More likely to be technically competent. Nature of funding that can help ensure timely completion.	Likely to be expensive. May not be sensitive to educational context. Difficult to ensure assessment sustainability. Possibly little national capacity enhancement.
Made up of a national team supported with some international TA	Can improve technical capacity of nationals. May ensure timely completion. May add credibility to the results.	Possibly difficult to coordinate work of national team members and TA. Might be difficult to ensure skill transfer to nationals.
Ministry team supported with national TA	Can ensure ministry support while obtaining national TA. Less expensive than international TA.	National TA that may lack the necessary technical capacity. Other potential disadvantages that are listed under ministry of education and that may apply.

Source: Authors.

responsible for the design of achievement tests, the analyses of results, and the general coordination of annual assessment activities.

It is worth reflecting on the wide variety of skills that are required to carry out a national assessment in deciding who should be given responsibility for the task. This issue is addressed in more detail in *Implementing a National Assessment of Educational Achievement* (book 3 in this series). A national assessment is fundamentally a team effort. The team should be flexible, willing to work under pressure and in a collaborative manner, and prepared to learn new assessment and technological approaches. The team leader should have strong managerial skills. He or she will be required to organize the staff, to coordinate and schedule activities, to support training, and to arrange and monitor finance. The team leader should be politically astute because he or she will need to report to an NSC and to be a liaison with national, regional, and, in some instances, district-level government bodies and representatives of stakeholders (such as teachers and religious bodies).

The team should have high-level implementation or operational skills. Tasks to be completed include organizing workshops for item writers and test administrators; arranging for printing and distribution of tests, questionnaires, and manuals; contacting schools; developing training materials; and collecting and recording data. A small dedicated team of test developers will be needed to analyze the curriculum, develop tables of specifications or a test blueprint, draft items, select items after pretesting or piloting, and advise on scoring. Following test administration, open-ended and multiple-choice questions have to be scored.

The team will require support from one or more people with statistical and analytical competence in selecting samples, in weighting data, in data input and file preparation, in item analysis of test data as well as general statistical analysis of the overall results, and in preparing data files for others (for example, academics and postgraduate students) to carry out secondary analyses. Many developing countries lack capacity in this last area, leading to situations in which data are collected but never adequately analyzed or reported.

The team should have the necessary personnel to draft and disseminate results, press releases, and focused pamphlets or newsletters.

It might also be reasonably expected to play a key role in organizing workshops for teachers and other education officials so they can discuss the importance of the results and the results' implications for teaching and learning.

Most members of the team may work part time and be employed as needed. This category could include item writers—especially practicing teachers with a good knowledge of the curriculum—and experts in sampling and statistical analysis. Team members might be recruited from outside the education sector. For example, a national census bureau can be a good source of sampling expertise. Computer personnel with relevant experience could help with data cleaning, and journalists could assist with drafting catchy press releases. Neither Cambodia nor Ethiopia employed full-time staff members to carry out its national assessment.

Responsibility for carrying out national assessment: Implementation agency (ministry of education, examination board, research agency, university).

WHO WILL ADMINISTER THE TESTS AND QUESTIONNAIRES?

National administrative traditions and perceptions of levels of trust, as well as sources of finance, tend to influence the selection of personnel responsible for administering tests and questionnaires in a national assessment. Practice varies. For example, some countries have used graduate students, while Zambia has involved school inspectors and ministry officials in test and questionnaire administration. Other countries have used experienced teachers drawn from nonparticipating schools or retired teachers. In the Maldives, a test administrator must be a staff member of a school located on an island other than the island where the targeted school is located.

Test administrators should be carefully selected. They should have good organizational skills, have experience of working in schools, and be committed to following test and questionnaire guidelines precisely. Ideally, they should have classroom experience, speak in the same language and accent as the students, and have an authoritative but nonthreatening manner. Book 3 of this series, *Implementing a National*

Assessment of Educational Achievement, considers the advantages and disadvantages of having teachers, inspectors, teacher trainers, examination board personnel, and university students as administrators.

Although the use of teachers of students who are participating in the national assessment as test administrators may appear administratively convenient and very cost-effective, it is, for a variety of reasons, rarely done. Some teachers might feel that their teaching effectiveness is being evaluated. Some may find it difficult to desist from their normal practice of trying to help students and might not be able to adjust to the formal testing approach. Some may make copies of tests or test items, thus ruling out the possibility of using those items in future national assessments. Having teachers administer tests to their own students might also diminish the public perception of the trustworthiness of the assessment results.

Responsibility for administering tests and questionnaires: Implementation agency

WHAT POPULATION WILL BE ASSESSED?

As the term is usually understood, national assessments refer to surveys carried out in education systems. This connotation, however, was not always the case. When the first national assessment was carried out in the United States (in 1969), out-of-school populations (17- and 18-year-olds and young adults 26–35 years of age), as well as school-going populations, were assessed (in citizenship, reading, and science). The assessment of the out-of-school populations was discontinued, however, because of cost (Jones 2003). Subsequent surveys of adult literacy were carried out independent of national assessments.

The issue of assessing younger out-of-school children is more relevant in many developing countries than in the United States because many children of school-going age do not attend school. Obviously, the achievements (or lack of them) of those children are of interest to policy makers and politicians and may have particular relevance for the nonformal education sector. Their inclusion in a conventional national assessment is, however, difficult to envisage. Although particular groups of out-of-school youth might be assessed

using national assessment tests in a separate study, methods of assessment and sampling procedures generally would be very different, and the varying circumstances of such children (for example, special needs, socioeconomic disadvantage, or distance from school) would have to be taken into account.

As far as school-going children are concerned, policy makers want information about their knowledge and skills at selected points in their educational careers. A decision has to be made about whether populations are defined on the basis of age or grade or, indeed, by a combination of age and grade. In countries where students vary widely in the age at which they enter school, and in which policies of nonpromotion are in operation, students of similar age will not be concentrated in the same grade. In this situation, a strong argument can be made for targeting grade level rather than age.

The grade to be assessed should normally be dictated by the information needs of the ministry of education. If, for example, the ministry is interested in finding out about the learning achievement levels of students completing primary school, it might request that a national assessment be carried out toward the end of the last year of primary school (fifth or sixth grade in many countries). The ministry could also request a national assessment in third or fourth grade if it needed data on how students are performing midway through the basic education cycle. This information could then be used to introduce remedial measures (such as in-service courses for teachers) to address problems with specific aspects of the curriculum identified in the assessment.

Target grades for national assessments have varied from country to country. In the United States, student achievement levels are assessed at grades 4, 8, and 12; in Colombia, achievement is assessed at grades 3, 5, 7, and 9; in Uruguay, at preschool and at grades 1, 2, and 6; and in Sri Lanka, at grades 4, 8, and 10. In anglophone Africa, a regional consortium of education systems, the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), assessed grade 6 students. Countries in the francophone African consortium Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (Conférence des Ministres de l'Education des Pays ayant le Français en Partage) assessed students in grades 2 and 5.

Sometimes pragmatic considerations dictate grade selection. The Nigerian Federal Ministry of Education decided to assess students in grade 4 because testing at any lower level would have required translation of tests into many local languages. More senior grades were not considered suitable because students and teachers would be focused on secondary-school entrance examinations.

Relatively few countries conduct large-scale assessments in grades 1 to 3. Students at that level might not be able to follow instructions or to cope with the cognitive tasks of the assessment or with the challenge of completing multiple-choice tests. A Jamaican study noted that a sizable number of grade 1 students were unable to recognize the letters of the alphabet (Lockheed and Harris 2005). Nevertheless, we should bear in mind that because information about early student learning patterns may be critical to reform efforts, alternative procedures to monitor those patterns should be in place.

Responsibility for selecting population to be assessed: Ministry of education and NSC

WILL A WHOLE POPULATION OR A SAMPLE BE ASSESSED?

Most national and all regional and international studies use sample-based approaches in determining national achievement levels. Some national assessments have used both census- and sample-based approaches (for example, Costa Rica, Cuba, France, Honduras, Jordan, Mexico, and Uruguay), whereas most subnational assessments collect census data (for example, Minas Gerais, Parana, and São Paulo, Brazil; Bogotá, Colombia; and Aguascalientes, Mexico) (see Crespo, Soares, and deMello e Souza 2000). Several factors favor the use of a sample if the objective is to obtain information for policy purposes on the functioning of the education system as a whole. Those factors include (a) reduced costs in test administration and in cleaning and managing data, (b) less time required for analysis and reporting, and (c) greater accuracy because of the possibility of providing more intense supervision of fieldwork and data preparation (Ross 1987).

As noted in chapter 3, the purpose of an assessment is key in determining whether to test a sample or the entire population of targeted

TABLE 4.2

Advantages and Disadvantages of Census-Based Assessment to Hold Schools Accountable

Advantages	Disadvantages
Focuses on what are considered important aspects of education.	Tends to lead to neglect of subject areas that are not tested.
Highlights important aspects of individual subjects.	Tends to lead to neglect of aspects of subjects that are not tested (such as oral fluency in language).
Helps ensure that students reach an acceptable standard before promotion.	Has contributed to early dropout and nonpromotion.
Allows for direct comparisons of schools.	Leads to unfair ranking of schools where different social backgrounds are served and where results are not significantly different.
Builds public confidence in the performance of the system.	Has led to cheating during test administration and to subsequent doctoring of results.
Puts pressure on students to learn.	Tends to emphasize memorization and rote learning.
Results in some schools and students raising test performance levels.	Improved performance may be limited to a particular test and will not be evident on other tests of the same subject area.
Allows parents to judge the effectiveness of individual schools and teachers.	Leads to unfair assessment of effectiveness on the basis of test score performance rather than taking into account other established factors related to learning achievement.
Tends to be popular with politicians and media.	Seldom holds politicians accountable for failure to support delivery of educational resources.

Source: Authors.

students. On the one hand, the decision to involve an entire population may reflect an intention to foster school, teacher, or even student accountability. It facilitates the use of sanctions (incentives or penalties), the provision of feedback to individual schools on performance, and the publication of league tables, as well as the identification of schools with

the greatest need for assistance (for example, as in Chile and Mexico). On the other hand, the sample-based approach will permit the detection of problems only at the system level. It will not identify specific schools in need of support, although it can identify types or categories of schools (for example, small rural schools) that require attention. It can also identify problems relating to gender or ethnic equity.

An argument against the use of a sample-based approach is that because the assessment does not have high stakes attached to performance, some students will not be motivated to take the test seriously. That was not the case, however, in many countries—including South Africa—where some students were afraid that performance on the Trends in International Mathematics and Science Study (TIMSS) tests would count toward their official school results. It is interesting to note that cheating occurred during test administration, presumably because of the perception that relatively high stakes were attached to performance (see A.4 in appendix A).

Advantages and disadvantages of using a national assessment to hold schools, teachers, and sometimes students accountable are set out in table 4.2. The topics listed are derived for the most part from studies of the effects of high-stakes public examinations, not from a study of national assessments. Nevertheless, they should be relevant to census-based national assessments, at least to ones that act as surrogate public examinations (as in the United States and some Latin American countries).

Responsibility for deciding whether to use a sample or census: Ministry of education

WHAT WILL BE ASSESSED?

All national assessments measure cognitive outcomes of instruction or scholastic skills in the areas of language/literacy and mathematics/numeracy, a reflection of the importance of those outcomes for basic education. In some countries, knowledge of science and social studies is included in an assessment. Whatever the domain of the assessment, providing an appropriate framework is important, in the first instance for constructing assessment instruments and afterward for interpreting

results. The framework may be available in a curriculum document if, for example, the document provides expectations for learning that are clearly prioritized and put into operation. In most cases, however, such a framework will not be available, and those charged with the national assessment will have to construct it. In that task, close cooperation will be required between the assessment agency, those responsible for curricula, and other stakeholders.

Assessment frameworks attempt to clarify in detail what is being assessed in a large-scale assessment, how it is being assessed, and why it is being assessed (see Kirsch 2001). The aim of the framework is to make the assessment process and the assumptions behind it transparent, not just for test developers but also for a much larger audience, including teachers, curriculum personnel, and policy makers. The framework usually starts with a general definition or statement of purpose that guides the rationale for the assessment and that specifies what should be measured in terms of knowledge, skills, and other attributes. It then identifies and describes various performances or behaviors that will reveal those constructs by identifying a specific number of characteristic tasks or variables to be used in developing the assessment, and it indicates how those performances are to be used to assess student performance (Mullis and others 2006).

Many national assessments have been based on a content analysis at a particular grade level of what students are expected to have learned as a result of exposure to a prescribed or intended curriculum. Typically, this analysis is done in a matrix with cognitive behaviors on the horizontal axis and with themes or content areas on the vertical axis. Thus, the intersection of a cognitive behavior and content area will represent a learning objective. Cells may be weighted in terms of their importance.

Recent national (and international) assessments have drawn on research relating to the development in students of literary and numeracy skills that may or may not be represented in national curricula. For example, in the International Association for the Evaluation of Educational Achievement (IEA) *Framework and Specifications* document for the Progress in International Reading Literacy Study (PIRLS) 2006, reading literacy is defined as "the ability to understand and use those written language forms required by society and/or

valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment" (Mullis and others 2006, 3). From this definition it is evident that reading is much more than decoding text or getting the meaning of a passage or poem. PIRLS further clarified what it proposed to measure by indicating the process and tasks to be assessed and the percentages of test items devoted to each (table 4.3).

The framework document specified that the assessment would use test booklets with five literary and five informational passages and that each passage would be followed by 12 questions, half of which

TABLE 4.3

PIRLS Reading Comprehension Processes

Comprehension processes	Examples of tasks	Items
Focus on and retrieve explicitly stated information	Looking for specific ideas. Finding definitions or phrases Identifying the setting for a story (for example, time, place). Finding topic sentence or main idea (explicitly stated).	20%
Make straightforward inferences	Inferring that one event caused another. Identifying generalizations in text. Describing the relationship between characters. Determining the referent of a pronoun.	30%
Interpret and integrate ideas and information	Determining the overall message or theme. Contrasting text information. Inferring a story's mood or tone. Interpreting a real-world application of text information.	30%
Examine and evaluate content, language, and textual elements	Evaluating the likelihood that the events described could happen. Describing how the author devised a surprise ending. Judging the completeness or clarity of information in text. Determining the author's perspectives.	20%

Source: Campbell and others 2001; Mullis and others 2006.

would be multiple choice and half would be constructed response. It also indicated that because reading attitudes and behaviors were important for the development of a lifelong reading habit and were related to reading achievement, PIRLS would include items in the student questionnaire to assess student reading attitudes and behaviors. It justified its selection of students in the fourth year of formal schooling as the target population for the assessment on the basis that the fourth year represented the transition stage from learning to read to reading to learn.

In its assessment framework, PIRLS recognized two main purposes that students have for reading:

- Reading for literacy experience
- Reading to acquire and use information.

It also gave a detailed justification for the emphasis that PIRLS placed on finding out more about the environment and the context in which students learn to read. This emphasis led to the inclusion of questionnaire items on home characteristics that can encourage children to learn to read: literacy-related activities of parents, language spoken in the home, links between the home and the school, and students' out-of-school literacy activities. School-level items covered school resources that can directly or indirectly affect reading achievement. The framework document also justified assessing classroom variables, such as instructional approaches and the nature of teacher training.

A further alternative to basing an assessment instrument on curriculum-embedded expectations or prescriptions, which is feasible in the case of older students, is to build a test to reflect the knowledge and skills that students are likely to need and build on in adult life. The Programme for International Student Assessment (PISA) provided an example of this method when it set out to assess the "mathematical literacy" of 15-year-olds, defined as the "capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in works that meet the needs of the individual's life as a constructive, concerned, and reflective citizen" (OECD 2003, 24) (see B.3 in appendix B). Although this approach fitted well in an international study, given that

the alternative of devising an assessment instrument that would be equally appropriate to a variety of curricula is obviously problematic, it might also be used in a national assessment.

A few national assessments have collected information on affective outcomes (for example, student attitudes to school and student self-esteem). In Colombia, for example, students' attitudes to peace are assessed. Although those outcomes are very important, their measurement tends to be less reliable than the measurement of cognitive outcomes, and analyses based on them have proved difficult to interpret. In Chile, technical difficulties in measuring student values and attitudes to learning led to abandoning those areas (see A.7 in appendix A).

One large-scale assessment (Monitoring Learning Achievement) assessed "life skills," defined as students' knowledge of, and attitudes toward, health and nutrition, environment, civic responsibility, and science and technology (Chinapah 1997). While it is generally accepted that life skills are important and should be taught, there is considerable disagreement about their precise nature. Their measurement has also proven difficult.

Most national assessments collect information on student, school, and home factors that are considered relevant to student achievement (for example, student gender and educational history, including grade repetition; resources in schools, including the availability of textbooks; level of teacher education and qualifications; and socioeconomic status of students' families). The information is normally collected in questionnaires (and sometimes in interviews) administered to students, to teachers, to principal teachers, and sometimes to parents at the same time as the assessment instruments are administered.

Identification of contextual factors related to student achievement can help identify manipulable variables, that is, factors that can be altered by policy makers, such as regulations about the time allocated to curriculum areas, textbook provision, and class size. The contextual data collected in some national (and international) studies, however, cannot play this role because they do not adequately measure the conditions in which students live. Economic status, for example, may be based on a list of items that includes a car, a television set, and a water tap in a country where the majority of the population lives at least part

of the year on less than the equivalent of US$1 a day. Furthermore, despite the relevance of health status and nutritional status, no information may be obtained about them (Naumann 2005).

In some assessments, teachers' (as well as pupils') achievements have been assessed. In Vietnam (see A.2 in appendix A) and a number of African countries in the SACMEQ studies (see C.1 in appendix C), teachers were required to take the same test items as their students to gain some insight into teachers' levels of subject mastery. In Uganda, information was obtained on the extent to which teachers claimed to be familiar with key official curriculum documents.

Responsibility for deciding what will be assessed: Ministry of education, NSC, with input from implementation agency.

HOW WILL ACHIEVEMENT BE ASSESSED?

An instrument or instruments must be devised that will provide the information that the national assessment is meant to obtain. Because the purposes and proposed uses of national assessments vary, so too will the instruments used in the assessments and the ways results are reported.

Some national assessments present results in terms of the characteristics of the distribution of test scores—for example, the mean percentage of items that students answered correctly and the way scores were distributed around the mean. Or results might be scaled to an arbitrary mean (such as 500) and standard deviation (such as 100). Although these scores can be used to compare the performance of subgroups in the sample, they are limited in their use in a national assessment, primarily because they tell us little about students' level of subject matter knowledge or the actual skills that students have acquired.

To address this issue, and to make the results of an assessment more meaningful for stakeholders, an increasing number of national assessments seek to report results in a way that specifies what students know and do not know and that identifies strengths and weaknesses in their knowledge and skills. This approach involves matching student scores with descriptions of the tasks they are able to do (for example, "can

read at a specified level of comprehension" or "can carry out basic mathematical operations"). Performances may be categorized in various ways (for example, "satisfactory" or "unsatisfactory"; "basic," "proficient," or "advanced"), and the proportion of students achieving at each level determined. Matching student scores to performance levels is a complex task involving the judgment of curriculum experts and statistical analysts.

The way in which results will be described should be a consideration at the test development stage. Thus, test development might begin with specification of a framework in which expectations for learning are posited, following which test items are written to assess the extent to which students meet those expectations. If items do not meet certain criteria when tried out, however, including the extent to which they discriminate between students, they may not be included in the final assessment instrument. Care should be taken to ensure that important curriculum objectives are reflected in an assessment, even if no students in the trial provide evidence of achieving them.

Most national and international assessments rely to a considerable extent on the multiple-choice format in their instruments. Those items will often be supplemented by open-ended items that require the student to write a word, phrase, or sentence. Examples of multiple-choice and open-ended items are provided in box 4.2 and box 4.3, respectively.

In several national (for example, the U.S. NAEP and Ireland's National Assessment of English Reading) and international assessments (for example, TIMSS and PISA), each student responds to only a fraction of the total number of items used in an assessment (see A.8 in appendix A; B.1 and B.3 in appendix B). This approach increases overall test coverage of the curriculum without placing too great a burden on individual students. It also allows the use of extended passages (for example, a short story or a newspaper article) in the assessment of reading comprehension. In other assessments, all students respond to the same set of items. Although some advantages are associated with having individual students respond to only a fraction of items, disadvantages also exist, particularly for countries beginning a national assessment program. Administration (for example, printing and distribution) is more complex, as is scoring and scaling of

BOX 4.2

Examples of Multiple-Choice Items

Subject: Geography

The river Volga is in

A. China

B. Germany

C. Russia

D. Sweden.

Subject: Mathematics

A seal has to breathe if it is asleep. Martin observed a seal for one hour. At the start of this observation, the seal dived to the bottom of the sea and started to sleep. In eight minutes, it slowly floated to the surface and took a breath. In three minutes, it was back at the bottom of the sea again, and the whole process started over in a very regular way. After one hour, the seal was

A. at the bottom

B. on its way up

C. breathing

D. on its way down.

Source: Mathematics example: OECD 2007. Reproduced with permission.

BOX 4.3

Examples of Open-Ended Items

Subject: Language

TALL is the opposite of SMALL.

What is the opposite of

QUICK _____ DARK _____

HEAVY _____ OLD _____

Subject: Mathematics

Use your ruler to draw a rectangle with a perimeter of 20 centimeters. Label the width and the length.

scores, while analyses involving individual student or school data can be problematic (see Sofroniou and Kellaghan 2004).

The issue of language of assessment is generally accorded less attention than it deserves. It is associated with two problems. First, although in many countries large minority (and sometimes majority) groups are present for whom the language of instruction is not their mother tongue, students are usually assessed in the language of instruction. In Uganda, for example, the vast majority of students take tests in their second language (see A.9 in appendix A). Poor performance on tests is attributed to this practice, as are the generally poor scholastic progress of students and early dropout rates from school (Naumann 2005).

A second problem relating to language arises if the instruments of the assessment need to be translated into one or more languages. If comparisons are to be made between performances assessed in different languages, analysis must take into account the possibility that differences that may emerge may be attributable to language-related differences in the difficulty of assessment tasks. The issue is partly addressed by changing words. For example, in an international assessment carried out in South Africa, words such as "gasoline" ("petrol") and "flashlight" ("torch") were changed. Ghana replaced the word "snow" with "rain." If language differences co-vary with cultural and economic factors, the problem is compounded because it may be difficult to ensure the equivalence of the way questions are phrased and the cultural appropriateness of content in all language versions of a test. For example, material that is context-appropriate for students in rural areas—covering hunting, the local marketplace, agricultural pursuits, and local games—might be unfamiliar to students in urban areas.

Whatever the details of the method of assessment, the assessment needs to provide valid and reliable information. Validity has several facets, including the adequacy of an assessment instrument to sample and represent the construct (for example, reading literacy) or the curriculum area (for example, social studies) identified in the assessment framework. The judgment of curriculum specialists is important here. Furthermore, the assessment instrument should measure only what it is designed to measure. For example, a test of mathematics or science should assess students' knowledge and skills in those areas, not their

competence in language. The reliability of assessment procedures in national assessments usually involves estimating the extent to which individual items in a test assess the overall construct the test is designed to measure and, in the case of open-ended items, the extent to which two or more markers agree in their scoring.

Responsibility for deciding how achievement will be assessed: Implementation agency.

HOW FREQUENTLY WILL ASSESSMENTS BE CARRIED OUT?

The frequency with which a national assessment is carried out varies from country to country, ranging from every year to every 10 years. A temptation may exist to assess achievement in the same curriculum areas and in the same population every year, but this frequency is unnecessary, as well as very expensive, if the aim is to monitor national standards. In the United States, reading and mathematics are assessed every second year and other subjects less frequently. The international assessment of reading literacy (PIRLS) had a five-year span between the first and second administration (2001–06). In Japan, achievement in core curriculum areas was assessed every 10 years to guide curriculum and textbook revision (Ishino 1995).

If the aim of an assessment is to hold teachers, schools, and even students accountable for their learning, testing may be carried out every year. Furthermore, because such an assessment focuses on the performance of individuals, as well as performance at the system level, all (or most) students in the education system will be assessed. This system has been operated in Chile and in England.

If the purpose of an assessment is only to provide information on the performance of the system as a whole, however, an assessment of a sample of students in a particular curriculum area every three to five years would seem adequate. Because education systems do not change rapidly, more frequent assessments would be unlikely to register change. Overfrequent assessments would more than likely limit the impact of the results, as well as incur unnecessary costs.

Responsibility for deciding frequency of assessment: Ministry of education

HOW SHOULD STUDENT ACHIEVEMENT BE REPORTED?

Although policy makers probably prefer summary statistics, evidence on the multidimensionality of achievement suggests that a single index of performance, such as a total test score, may obscure important information. An alternative approach is to provide differentiated information that reflects strengths and weaknesses in a country's curriculum. The information would be even more valuable if it distinguished between students' knowledge of basic facts and skills and their deeper or higher-order understanding.

A variety of procedures have been used to describe student achievements in national assessments, which reflect the richness of the data that an assessment can provide (see book 5 in this series, *Reporting and Using Results from a National Assessment of Educational Achievement*). The selection of one or more procedures should be guided by the information needs of the ministry of education and other stakeholders.

Item-Level Information

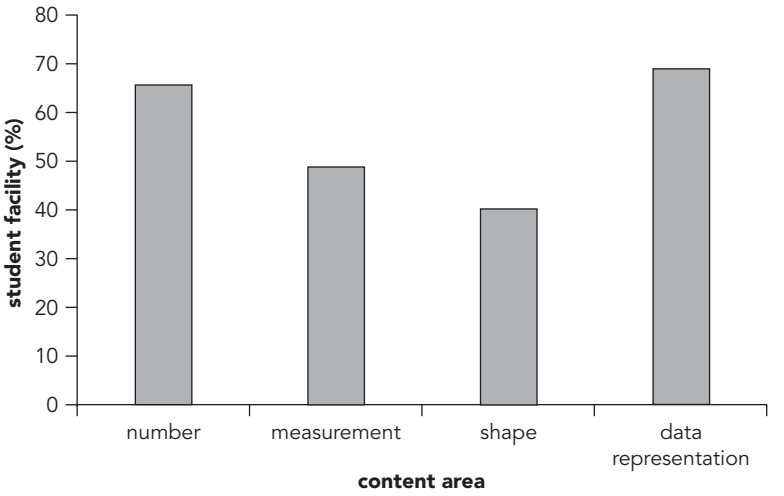
This information involves little more than simply reporting the percentage of students answering individual items correctly. A national assessment might reveal that the majority of its students performed poorly on a mathematics item involving the use of indices, or that virtually all students were able to associate simple words with pictures. In Ghana, for example, only 1 percent of students correctly answered a question on light refraction in TIMSS (Ghana, Ministry of Education, Youth, and Sports 2004). This kind of information, while too detailed for national policy making, is likely to be of interest to curriculum personnel, teacher trainers, and possibly textbook authors.

Performance in Curriculum Domains

Items can be grouped into curriculum units or domains, and test scores can be reported in terms of performance in each domain. Reading items, for example, have been classified by ability to retrieve information from a text, to make inferences from a text, to interpret

FIGURE 4.1

Mean Percentage Correct Scores for Students' Mathematics Performance, by Content Area, Lesotho



Source: Lesotho, Examinations Council of Lesotho and National Curriculum Development Centre 2006.

and integrate information, and to examine and evaluate text information (Eivers and others 2005). Figure 4.1 illustrates how Lesotho reported mathematics performance by content area.

Performance Standards

Performance on national and international assessments describes how well students perform on the test to achieve a "basic," "proficient," or "advanced" level in a curriculum area. The number of levels may vary (see A.2 in appendix A for a description of six levels of reading proficiency used in a national assessment in Vietnam, and see C.1 in appendix C for eight reading levels and eight mathematics skill levels used in SACMEQ). The selection of cutoff points between levels involves the use of statistical data and subjective judgment.

Mastery Standard

Mastery levels can be based on an overall test score (for example, correctly answering a specified percentage of test items). In Sri Lanka,

the mastery level for a grade 4 national assessment was set at 80 percent. Fewer than 40 percent achieved that level in the students' first language or in mathematics, and fewer than 10 percent in English (Perera and others 2004). Mastery levels can also be based on achieving a certain performance level. In the United States, five levels of performance ("below basic," "basic," "proficient," "goal," and "advanced") are used in Connecticut. The "goal" level is regarded as a challenging but reasonable level of expectation for students and is accepted as the mastery level. The data in table 4.4 show that well over half the students in grades 3 and 4 achieved the "goal" or "mastery" level in all three curriculum areas.

Responsibility for deciding how student achievement is reported: Implementation agency with input from NSC

WHAT KINDS OF STATISTICAL ANALYSES SHOULD BE CARRIED OUT?

Some analyses will be dictated by the policy questions that prompted the assessment in the first instance. Most national assessments provide evidence on achievement by gender, region, urban or rural location, ethnic or language group membership, and type of institution attended (public or private). Some assessments also provide data on the quality of school facilities (for example, Kenya). Analyses involving those variables are relatively straightforward and are intuitively meaningful to policy makers and politicians. They do not, however, adequately represent the complexity of the data. More complex forms of analysis are required if we are, for example, to throw light on the school and background factors that contribute to achievement. Examples of the use of complex statistical procedures are found in the description of the Vietnamese national assessment (see A.2 in appendix A).

The limitations of analyses and problems in inferring causation from studies in which data are collected at the same time on achievement and other variables should be recognized. Although it is difficult, sometimes impossible, to disentangle the effects of community, home, and school factors on students' learning, this complexity has not deterred

TABLE 4.4

Percentage Achieving Goal or Mastery Level by Grade, Connecticut, 2006

	Mathematics		Reading		Writing	
Grade	At or above goal (%)	At or above advanced (%)	At or above goal (%)	At or above advanced (%)	At or above goal (%)	At or above advanced (%)
3	56	22	54	17	61	22
4	59	22	58	16	63	22

Source: Connecticut Department of Education 2006.

some investigations from causally interpreting data collected in national and international assessments.

Responsibility for deciding on methods of statistical analysis: Implementation agency.

HOW SHOULD THE RESULTS OF A NATIONAL ASSESSMENT BE COMMUNICATED AND USED?

If the results of a national assessment are to affect national education policy, they should be reported as soon as possible after the completion of data analysis. In the past, technical reports that featured a considerable amount of data tended to be the sole form of reporting. Some groups of users (for example, teachers in Chile; see A.7 in appendix A), however, considered those reports overtechnical. As a result, the requirement to provide other forms of reports is now increasingly recognized. Those alternatives include short summary reports that focus on the main findings for busy policy makers; press releases; special reports for radio and television; and separate reports for schools, teachers, curriculum developers, and teacher trainers. In some countries (for example, Sri Lanka), separate reports are prepared for each province. A report in Ethiopia was translated into four major languages. The information needs of stakeholders should determine the contents of additional reports.

The ministry of education should make adequate budgetary provision at the planning stage for report preparation and dissemination. In collaboration with the national steering committee, it should devise procedures to communicate the findings of national assessments to stakeholders. Appropriate strategies to communicate results should take into account the fact that users (whether administrators or teachers) vary greatly in their ability to understand and apply statistical information in their decision making. Obviously, there is no point in producing reports if the information they contain is not adequately disseminated. Thus, a dissemination strategy is also required so that relevant information reaches all stakeholders. The strategy should identify potential users (key institutions and individuals) and their levels of technical expertise.

National assessment results have been used to set benchmarks for monitoring learning achievement levels (for example, in Lesotho), reforming curricula, providing baseline data on the amount and quality of educational materials in schools (for example, in Vietnam), identifying correlates of achievement, and diagnosing aspects of the curriculum that are not being mastered by students. Uruguay, for instance, used its national assessment results to help prepare teacher guides and to identify the curriculum content and behavioral areas that subsequently helped direct a large-scale teacher in-service program (see A.3 in appendix A).

Book 5 in this series, *Reporting and Using Results from a National Assessment of Educational Achievement*, has an extensive section on report writing and the use of national assessment results.

Responsibility for communicating and using national assessment results: Implementation agency, ministry of education, NSC, teacher training providers, curriculum authority, teachers.

WHAT ARE THE COST COMPONENTS OF A NATIONAL ASSESSMENT?

The cost of a national assessment will vary greatly from one country to another, depending on the salary levels of personnel and the cost of services. Within a country, cost will also vary, depending on some or all of the following factors (Ilon 1996).

- *Implementing agency.* Costs will vary depending on whether the agency has the necessary facilities and expertise or needs to upgrade or employ full-time or part-time consultants. The cost of providing facilities and equipment, including computers and software, also needs to be taken into account.
- *Instrument content and construction.* Options for the selection of the content and form of assessment should be considered in terms of cost, as well as other factors, such as validity and ease of administration. Multiple-choice items are more expensive to construct than open-ended items but are usually less expensive to score. The cost of translating tests, questionnaires, and manuals and of training item writers also needs to be considered.

- *Numbers of participating schools and students.* A census-based assessment will obviously be more expensive than a sample-based one. Costs increase if reliable data are required for sectors of the system (for example, states or provinces). Targeting an age level is likely to be more expensive than targeting a grade level because students of any particular age may be spread over a number of grades, requiring additional assessment material and testing sessions.
- *Administration*. Data collection tends to be the most expensive component of a national assessment. It involves obtaining information from schools in advance of the assessment; designing, printing, packaging, and dispatching test materials and questionnaires; and establishing a system to administer instruments. Factors that contribute to overall cost include (a) the number of schools and students that participate, (b) travel, (c) difficulty in gaining access to schools, (d) accommodation for enumerators (if needed), and (e) the collection and return of completed tests and questionnaires.
- *Scoring, data management, and data entry.* Costs will vary according to the number of participating schools, students, teachers, and parents; the number of open-ended items; whether items are hand or machine scored; the number of inter-rater reliability studies; and the quality of test administration and scoring.
- *Analysis.* Analytic costs will depend on the type of assessment procedures used and the availability of technology for scoring and analysis. Although machine scoring is normally considered to be cheaper than hand scoring, this reduced cost may not be the case in a country where technology costs are high and labor costs are low.
- *Reporting.* Costing should take account of the fact that different versions of a report will be required for policy makers, teachers, and the general public and of the nature and extent of the report dissemination strategy.
- *Follow-up activities.* Budgetary provision may have to be made for activities such as in-service teacher training that is based on the findings of the national assessment, briefings for curriculum bodies, and secondary analyses of the data. Provision may also have to be made to address skill shortages in key professional areas (for example,

statistical analysis). Budgetary provision should be made for likely salary increases over the life of the assessment (normally two to three years), for inflation, and for unexpected events (contingencies).

Some national assessments have not achieved their basic objectives because the budget was inadequate. Although the overall budget is the responsibility of the ministry of education, people with expertise in costing and with large-scale data projects should participate in the budgetary discussions. Ministry officials who are unfamiliar with large-scale data projects are unlikely to appreciate the need to budget for activities such as pilot-testing and data cleaning.

Figures for the U.S. NAEP provide a rough guide to costing: data collection (30 percent), instrument development (15 percent), data analysis (15 percent), reporting and dissemination (15 percent), sampling (10 percent), data processing (10 percent), and governance (5 percent) (Ilon 1996). In some countries, where, for example, ministry or examination board officials carry out test administration as part of their normal duties, separate budgetary provision may not be made for some activities. Costs and wages will vary depending on national economic conditions. In Cambodia (which is ranked outside the top 100 countries in the world in terms of gross national income), item writers were paid the equivalent of US$5 a day in 2006.

Countries with very limited resources may not find expending those resources on a national assessment advisable, especially when their education system is likely to have many unmet needs. If they do wish to engage in national assessment activity, they would be well advised to limit the number of curriculum areas assessed (perhaps to one, at one grade level) and to seek technical assistance and the support of donors.

In considering costs, it is well to bear in mind that the cost of accountability programs in general—and of national assessments in particular—is very small compared to the cost of other educational programs (see Hoxby 2002). The cost of *not* carrying out an assessment—of not finding out what is working and what is not working in the education system—is likely to be much greater than the cost of an assessment. Book 3 of this series, *Implementing a National Assessment of Educational Achievement*, discusses issues relating to costing a national assessment.

Responsibility for estimating the component costs of a national assessment: Ministry of education with consultant input.

SUMMARY OF DECISIONS

Table 4.5 identifies the agencies with primary responsibility for decisions relating to the 12 components of a national assessment that are discussed in this chapter.

TABLE 4.5

Bodies with Primary Responsibility for Decisions in a National Assessment

Decision	Primary responsibility			
	Ministry of education	National Steering Committee	Agency	Other
Give policy guidance	•			
Carry out national assessment			•	
Administer tests and questionnaires			•	
Choose population to be assessed	•	•		
Determine sample or population	•			
Decide what to assess	•	•	•	
Decide how achievement is assessed			•	
Determine frequency of assessment	•			
Select methods of reporting		•	•	
Determine statistical procedures			•	
Identify methods of communicating and using results	•	•	•	•
Estimate cost components	•			•

Source: Authors.

CHAPTER 5

ISSUES IN THE DESIGN, IMPLEMENTATION, ANALYSIS, REPORTING, AND USE OF A NATIONAL ASSESSMENT

In this chapter, we identify a number of issues that are relevant to the confidence that stakeholders can have in the results of a national assessment. For five components of national assessment activity (design, implementation, data analysis, report writing, and dissemination and use of findings), we suggest a number of activities that will enhance confidence, which, in turn, should contribute to the optimum use of findings. For each component, we also identify common errors that have been made in national assessments and that evaluators should be aware of and should avoid.

DESIGN

The design of the assessment sets out the broad parameters of the exercise: the achievements to be assessed, the grade or age level at which students will be assessed, the policy issues to be addressed, and whether the assessment will involve the whole population or a sample of students.

Recommended Activities

- Involve senior policy makers from the outset to ensure political support and to help frame the assessment design.
- Determine and address the information needs of policy makers when selecting aspects of the curriculum, grade levels, and population subgroups (for example, by region or by gender) to be assessed.
- Obtain teacher support by involving teacher representatives in assessment-related policy decisions.
- Be aware that attaching high stakes to students' performance may lead to teacher opposition and to a narrowing of the effective curriculum as teachers focus their teaching on what is assessed.

Common Errors

- Failure to make adequate financial provision for key aspects of a national assessment, including report writing and dissemination.
- Failure to set up a national steering committee and to use it as a source of information and guidance during the course of the national assessment.
- Failure to gain government commitment to the process of national assessment, which is reflected in (a) a failure to identify key policy issues to be addressed at the design stage of the assessment, (b) the absence of a national steering committee, or (c) separate national assessments being carried out at the same time (often supported by external donors).
- Failure to involve key stakeholders (for example, teachers' representatives or teacher trainers) in planning the national assessment.
- Omission of a subgroup from the population assessed that is likely to seriously bias the results of the assessment (for example, students in private schools or students in small schools).
- Setting unrealistic test score targets (for example, 25 percent increase in scores over a four-year period).
- Allowing inadequate time for test development.

IMPLEMENTATION

Implementation covers a vast range of activities, from the development of appropriate assessment instruments, to the selection of the students who will respond to the instruments, to the administration of the instruments in schools.

Recommended Activities

- Describe in detail the content and cognitive skills of achievement and the background variables to be assessed.
- Entrust test development to personnel who are familiar with both curriculum standards and learning levels of students (especially practicing teachers).
- Use assessment instruments that adequately assess the knowledge and skills about which information is required and that will provide information on subdomains of knowledge or skills (for example, problem solving) rather than just an overall score.
- Develop clear and unambiguous test and questionnaire items, and present them in a clear and attractive manner.
- Ensure that adequate procedures are in place to assess the equivalence of language versions if translation of instruments is necessary.
- Pilot-test items, questionnaires, and manuals.
- Review items to identify ambiguities and possible bias relating to student characteristics (for example, gender, location, or ethnic group membership), and revise or delete if necessary.
- Proofread all materials carefully.
- Establish procedures to ensure the security of all national assessment materials (for example, tests and questionnaires) throughout the whole assessment process, so that materials do not fall into the hands of unauthorized people.
- Secure the services of a person or unit with sampling expertise.
- Specify the defined target population (the population from which a sample will actually be drawn—that is, the sampling frame) and the excluded population (for example, elements of the population

that are too difficult to reach or that would not be able to respond to the instrument). Precise data on excluded populations should be provided.
- Ensure that the proposed sample is representative and is of sufficient size to provide information on populations of interest with an acceptable level of error.
- Select members of the sample from the sampling frame according to known probabilities of selection.
- Follow a standard procedure when administering tests and questionnaires. Prepare an administration manual.
- Ensure that test administrators are thoroughly familiar with the contents of tests, questionnaires, and manuals and with administrative procedures.
- Prepare and implement a quality assurance mechanism to cover, among other things, test validity, sampling, printing, test administration, and data preparation.

Common Errors

- Assigning test development tasks to people who are unfamiliar with the likely levels of student performance (for example, academics), resulting in tests that are too difficult.
- Representing curriculum inadequately in tests, as indicated, for example, in failure to include important aspects of the curriculum.
- Failing to pilot-test items or pilot-testing on an unrepresentative sample of the population.
- Using an insufficient number of test items in the final version of the test.
- Failing to give a clear definition of the construct being assessed (for example, reading).
- Including an insufficient number of sample items for students who are unfamiliar with the testing format.
- Not encouraging students to seek clarification from the test supervisor before taking the test.
- Failing to give adequate notification to printers of tests, questionnaires, and manuals.
- Paying insufficient attention to proofreading tests, questionnaires, and administration manuals.

- Using inadequate or out-of-date national data on pupils and school numbers for sampling.
- Failing to carry out proper sampling procedures, including selecting a predetermined percentage of schools (for example, 5 percent).
- Providing inadequate training to test and questionnaire administrators.
- Allowing outside intervention (for example, principal sitting in the classroom) during test administration.
- Allowing students to sit close to each other during the assessment (encourages copying).
- Failing to establish a tradition of working outside normal work hours, if needed, to complete key tasks on time.

ANALYSIS

Statistical analyses organize, summarize, and interpret the data collected in schools. They should address the policy issues identified in the design of the national assessment.

Recommended Activities

- Secure competent statistical services.
- Prepare a codebook with specific directions for preparing data for analysis.
- Check and clean data to remove errors (for example, relating to numbers, out-of-range scores, and mismatches between data collected at different levels).
- Calculate sampling errors, taking into account complexities in the sample, such as stratification and clustering.
- Weight data so that the contribution of the various sectors of the sample to aggregate achievement scores reflects their proportions in the target population.
- Identify the percentage of students who met defined acceptable levels or standards.
- Analyze assessment data to identify factors that might account for variation in student achievement levels to help inform policy making.

- Analyze results by curriculum domain. Provide information on the subdomains of a curriculum area (for example, aspects of reading, mathematics).
- Recognize that a variety of measurement, curricular, and social factors may account for student performance.

Common Errors

- Using inappropriate statistical analyses, including failing to weight sample data in the analysis.
- Basing results on small numbers (for example, a minority of sampled teachers who might have responded to a particular question).
- Contrasting student performance in different curriculum areas, and claiming that students are doing better in one area on the basis of mean score differences.
- Failing to emphasize the arbitrary nature of selected test score cutoff points (for example, mastery versus nonmastery, pass versus fail), dichotomizing results, and failing to recognize the wide range of test scores in a group.
- Not reporting standard errors associated with individual statistics.
- Computing and publicizing school rankings on the basis of achievement test results without taking into account key contextual factors that contribute to the ranking. Different rankings emerge when school performances are compared using unadjusted performance scores, scores adjusted for contextual factors (for example, the percentage of students from poor socioeconomic backgrounds), and scores adjusted for earlier achievement.
- Inferring causation where it might not be justified (for example, attributing differences in learning achievement to one variable, such as private school administration or class size).
- Comparing test results over two time periods even though nonequivalent test items were used.
- Comparing test results over two time periods without reporting the extent to which important background conditions (for example, curriculum, enrollment, household income, or level of civil strife) might have changed in the interim. Although most education-related variables tend not to change rapidly over a short time (for example,

three to four years), some countries have introduced policies that have resulted in major changes in enrollment. Following the abolition of school fees, for example, the number of students enrolling in schools increased greatly in Malawi and Uganda.
- Limiting analysis in the main to a listing of mean scores of geographical or administrative regions.

REPORT WRITING

There is little point in carrying out a national assessment unless the findings are clearly reported with the needs of various stakeholders in mind.

Recommended Activities

- Prepare reports in a timely manner with the needs of clients in mind, and present them in a format that is readily understood by interested parties, especially those in a position to make decisions.
- Report results by gender and region, if sample design permits.
- Provide adequate information in the report or in a technical manual to allow for replication of the assessment.

Common Errors

- Writing overly technical reports.
- Failing to highlight a few main findings.
- Making recommendations in relation to a specific variable even though the analysis questioned the validity of the data on that variable.
- Failing to relate assessment results to curriculum, textbook, and teacher training issues.
- Not acknowledging that factors outside the control of the teacher and the school contribute to test score performance.
- Failing to recognize that differences between mean scores may not be statistically significant.
- Producing the report too late to influence relevant policy decisions.

- Doing an overextensive review of literature in the assessment report.
- Failing to publicize the key relevant messages of the report for separate stakeholder audiences.

DISSEMINATION AND USE OF FINDINGS

It is important that the results of national assessments are not left on policy makers' shelves but are communicated in appropriate language to all who can affect the quality of students' learning.

Recommended Activities

- Provide results to stakeholders, especially key policy makers and managers.
- Use the results where appropriate for policy making and to improve teaching and curricula.

Common Errors

- Ignoring the results when it comes to policy making.
- Among key stakeholders (for example, teacher trainers or curriculum personnel), failing to consider the implications of the national assessment findings.
- Among the national assessment team, failing to reflect on lessons learned and to take note of those lessons in follow-up assessments.

CHAPTER 6

INTERNATIONAL ASSESSMENTS OF STUDENT ACHIEVEMENT

In this chapter, we describe international assessments of students' educational achievement because they are used in many countries to provide data for a national assessment. First, we outline the main features of international assessments in terms of how they are similar to and differ from national assessments. Next, we describe growth in international assessment activity. Then the chapter identifies advantages of international assessments as well as problems associated with these assessments.

An international assessment of student achievement is similar in many ways to a national assessment. Both exercises use similar procedures (in instrument construction, sampling, scoring, and analysis). They also may have similar purposes: (a) to determine how well students are learning in the education system; (b) to identify particular strengths and weaknesses in the knowledge and skills that students have acquired; (c) to compare the achievements of subgroups in the population (for example, defined in terms of gender or location); or (d) to determine the relationship between student achievement and a variety of characteristics of the school learning environment and of homes and communities. Furthermore, both exercises may attempt to establish whether student achievements change over

time (Kellaghan and Greaney 2004). In practice, however, why a country decides to participate in an international assessment is not always clear (Ferrer 2006).

The main advantage of an international assessment compared to a national assessment is that the former has as an objective to provide policy makers, educators, and the general public with information about their education system in relation to one or more other systems (Beaton and others 1999; Husén 1973; Postlethwaite 2004). This information is assumed to put pressure on policy makers and politicians to improve services. Furthermore, it is hoped that the information will contribute to a greater understanding of the factors (that vary from country to country) that contribute to differences in student achievement.

The curriculum areas that have attracted the largest participation rates in international studies over the years are reading comprehension, mathematics, and science. Studies have been carried out at primary- and secondary-school levels. Usually, a combination of grade and age is used to determine who will participate (for example, students in two adjacent grades that contain the largest proportions of 9-year-olds and 13-year-olds; students in the grade levels containing most 9-year-olds and most 14-year-olds; the upper of two adjacent grades with the most 9-year-olds). In yet another international study, students of a particular age were selected (15-year-olds).

The results of international assessments such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) and regional assessments can and have been used to prepare separate national reports on country-level performance. International databases can be accessed to carry out such analyses.

Countries vary considerably in the extent to which they rely on international and national assessment results for policy making. Many industrial countries conduct their own national assessments, as well as participating in international assessments. The United States has its own National Assessment of Educational Progress for grades 4, 8, and 12; it also participates in international assessments of achievement. Some industrial countries have participated in international assessments but do not conduct national assessments (for example, the Russian Federation and Germany). Similarly, some developing countries

have used international assessments to provide their sole form of national assessment (Braun and Kanjee 2007). Many of the world's poorest countries have not taken part in international assessments or carried out national assessments, although the situation has changed in recent years.

GROWTH IN INTERNATIONAL ASSESSMENT ACTIVITY

International assessment activity began when a group of researchers met in 1958 to consider the possibility of undertaking a study of measured outcomes and their determinants within and between systems of education (Husén and Postlethwaite 1996). Since then, more than 60 countries have participated in international studies of achievement in one or more of a variety of curriculum areas: reading, mathematics, science, writing, literature, foreign languages, civic education, and computer literacy. The best-known international assessments are TIMSS (see B.1 in appendix B) and the Progress in International Reading Literacy Study (PIRLS) (see B.2 in appendix B) of the International Association for the Evaluation of Educational Achievement (IEA) and PISA (see B.3 in appendix B) of the Organisation for Economic Co-operation and Development (OECD). Regional assessments in reading and mathematics have been carried out in southern and eastern Africa (see C.1 in appendix C), in francophone Africa (see C.2 in appendix C), and in Latin America (see C.3 in appendix C). A number of features on which TIMSS and PISA differ are presented in table 6.1 (see also B.1 and B.3 in appendix B).

The number of countries participating in international studies has increased over the years. While typically fewer than 20 countries participated up to the 1980s, the IEA Reading Literacy Study attracted 32 countries in 1991. In 2003, 52 countries participated in TIMSS and 41 in PISA (30 member states of the OECD and 11 "partner" countries). Furthermore, international studies in recent years have accorded a major focus to monitoring performance over time. All three major current international assessments (TIMSS, PIRLS, and PISA) are administered on a cyclical basis and are now described as "trend" studies.

TABLE 6.1
Comparison of TIMSS and PISA

	TIMSS 2003	PISA 2003
Purposes	To provide comparative evidence on the extent to which students have mastered official school curriculum content in mathematics and science, which is common across a range of countries. To monitor changes in achievement levels over time. To monitor students' attitudes toward mathematics and science. To examine the relationship between a range of instructional and school factors and achievement. (Reading is covered in separate PIRLS assessment.)	To provide comparative evidence on the "yield" of the school system in the principal industrial countries, and to assess whether students can apply their knowledge and competencies in reading, mathematics, and science to real-world situations. To monitor changes in achievement levels and equity of learning outcomes over time. To monitor student approaches to learning and attitudes to mathematics, science, and reading. To provide a database for policy development.
Framework	Developed by content experts from some participating countries.	Developed by content experts from some participating countries.
Target population	Grades 4 and 8.	15-year-olds.
Curriculum appropriateness	Designed to assess official curriculum organized around recognized curriculum areas common to participating countries.	Designed to cover knowledge acquired both in school and out of school, defined in terms of overarching ideas and competencies applied to personal, educational, occupational, public, and scientific situations.

	TIMSS 2003	PISA 2003	
Item content differences (mathematics, grade 8)	Grade 8, item distribution: • Number, 30% • Algebra, 25% • Data, 15% • Geometry, 15% • Measurement, 15%	Mathematics, overarching ideas: • Quantity • Space and shape • Change and relationships • Uncertainty	Item distribution: • Number, 31.8% • Geometry, 21.2% • Statistics, 21.2% • Functions, 10.6% • Discrete math, 5.9% • Probability, 5.9% • Algebra, 3.5%
Cognitive processes	Grade 8: • Solving routine problems 40% • Using concepts 20% • Knowing facts and procedures 15% • Reasoning 25%	Item distribution: • Connection, 47% • Reproduction, 31% • Reflection, 22%	
Item types (mathematics)	About two-thirds being multiple-choice items, with the remainder being constructed-response or open-ended items.	About one-third being multiple-choice items, with the remainder generally being closed (one possible correct response) or open (more than one possible correct response) constructed-response items.	
Frequency	Every four years: equal emphasis on mathematics and science in each cycle.	Every three years: extensive coverage of one domain (subject) every nine years (reading in 2000, mathematics in 2003, and science in 2006), plus less extensive coverage of the other two every three years.	
Geographical coverage	48 countries: 20 high-income, 26 middle-income, and 2 low-income countries.	30 OECD countries as well as 11 other countries.	
Analysis	Four benchmark levels and a mean score, which are based on all participating countries.	Seven mathematics proficiency levels and a mean score, which are based on OECD countries.	

Source: TIMSS and PISA frameworks; U.S. National Center for Education Statistics n.d.; World Development Indicators database.

Participation by nonindustrial countries in international studies has generally been low. Nevertheless, in line with the general increase in the number of countries that have taken part in international studies, the number of nonindustrial countries has increased over the years. TIMSS attracted the largest numbers in 2003 (seven from Africa) and 2007 (six from Africa). As was the case generally in international studies, nonindustrial countries have shown a greater interest in taking part in studies of mathematics and reading than in studies of other curriculum areas.

Recent growth in participation in international studies can be attributed to globalization, to a movement in health and education to benchmark services against those in other countries, and to interest in global mandates. Some research evidence supports the view that educational quality (in particular those aspects of it represented by mathematics and science achievements) plays an important role in economic growth, though it is not entirely consistent across countries or over time (Coulombe, Tremblay, and Marchand 2004; Hanushek and Kimko 2000; Hanushek and Wössmann 2007; Ramirez and others 2006). Whatever the reason, education policy around the world has increasingly focused on the need to monitor aggregate student achievement in an international context.

ADVANTAGES OF INTERNATIONAL ASSESSMENTS

A variety of reasons have been proposed to encourage countries to participate in an international assessment of student achievement. Perhaps the most obvious is that international studies provide a comparative framework in which to assess student achievement and curricular provision in a country and to devise procedures to address perceived deficiencies (Štraus 2005). By comparing results from different countries, countries can use assessment results to help define what is achievable, how achievement is distributed, and what relationships exist between average achievement and its distribution. For example, can high average achievement coexist with narrow disparities in performance? Results from PISA suggest that it can.

Data on achievement provide only limited information. It has been argued that an advantage of international studies is that they can capitalize on the variability that exists across education systems, thereby broadening the range of conditions that can be studied beyond those operating in any one country (Husén 1973). On this basis, the analysis of data collected in these studies routinely considers associations between achievement and a wide range of contextual variables. The range of variables considered includes curriculum content, time spent on school work, teacher training, class size, and organization of the education system. Clearly, the value of international studies is enhanced to the extent that they provide researchers and policy makers with information that suggests hypotheses about the reasons students differ in their achievements from country to country. The studies also provide a basis for the evaluation of policy and practices.

International assessments have the potential to bring to light the concepts for understanding education that have been overlooked in a country (for example, in defining literacy or in conceptualizing curricula in terms of intention, implementation, and achievement; see, for example, Elley 2005). The assessments can also help identify and lead to questioning of assumptions that may be taken for granted (for example, the value of comprehensive compared to selective education, smaller class sizes being associated with higher achievement, or grade repetition benefiting students).

International studies are likely to attract the attention of the media and of a broad spectrum of stakeholders, such as politicians, policy makers, academics, teachers, and the public. Differences between countries in levels of achievement are obvious in the descriptive statistics that are provided in reports of the studies. Indeed, those differences are usually highlighted in "league tables" in which countries are ranked in terms of their mean level of achievement. The comparative data provided in these studies have more "shock value" than the results of a national assessment. Poor results can encourage debate, which, in turn, may provide politicians and other policy makers with a rationale for increased budgetary support for the education sector, particularly if poor results are associated with a lower level of expenditure on education.

An important feature of an international assessment is that it provides data that individual countries can use to carry out within-country analyses for what becomes, in effect, a national assessment report. This practice is followed by countries that participate in PISA (see B.3 in appendix B) and SACMEQ (see C.1 in appendix C). The practice is enhanced if, in addition to the data collected for the international study, data that relate to issues of specific interest or concern in individual countries are also collected.

Participation in international assessments has a number of practical advantages, particularly for countries that do not have the capacity in their universities to develop the kinds of skills needed in national assessments. First, a central agency may carry out national-level analyses that can be used in individual country reports. Second, studies may contribute to the development of local capacity in a variety of technical activities: sampling, defining achievements, developing tests, analyzing statistics, and writing reports. Third, staffing requirements and costs (for example, for instrument development, data cleaning, and analysis) may be lower than in national assessments because costs are shared with other countries.

A study of the effect of TIMSS on the teaching and learning of mathematics and science in participating countries provides evidence of the variety of activities that an international study can spawn (Robitaille, Beaton, and Plomp 2000):

- TIMSS results featured in parliamentary discussions about planned changes in education policy (Japan).
- The minister for education established a mathematics and science task force (New Zealand).
- The president directed that a "rescue package" be implemented to improve performance in science and mathematics (in which teacher training would receive particular attention) (the Philippines).
- National benchmarks were established in literacy and numeracy (Australia).
- Results contributed to the development of new educational standards in mathematics and science (Russian Federation).
- Results helped change the nature of public discussions in the field of education from opinion-based discussions to fact-based discussions (Switzerland).

- Results led to the development of instructional materials that are based on analysis of the common misconceptions and errors of students in their response to TIMSS tasks (Canada).
- Results accelerated changes in revision of curricula (Czech Republic; Singapore).
- TIMSS results were identified as one of a number of factors influencing policy changes in mathematics education (England).
- Committees were formed to revise mathematics and science curricula (Kuwait).
- New topics were added to the mathematics curriculum (Romania).
- New content was introduced to the mathematics and science curriculum relating to real-life situations (Spain).
- Results helped highlight the need to improve the balance between pure mathematics and mathematics in context (Sweden).
- TIMSS findings highlighted beliefs about gender differences and negative attitudes to science and mathematics and were used as a basis for curriculum reform and teachers' professional development (Republic of Korea).
- Results influenced the outcome of discussions about improving the organization of, and emphasis in, teacher education (Iceland).
- TIMSS results led to taking steps to strengthen teacher professional development in mathematics and science (Norway; the United States).
- A centralized examination system was established, partly in response to TIMSS results (Latvia).
- TIMSS findings influenced major changes in teaching, school and class organization, teacher education, and target-setting for schools (Scotland).
- TIMSS findings affected educational research, standards development, curriculum document development, teacher studies, mathematics and science teaching methodologies, and textbook development (Slovak Republic).

The results of analyses of PISA data have led to the following:

- Cast doubt on the value of extensive use of computers in the classroom to improve achievement.
- Highlighted the fact that level of national expenditure on education is not associated with achievement (among participating countries).

- Prompted general policy debate on education (Germany).
- Contributed to the development of the secondary-school science curriculum (Ireland).
- Emphasized the complexity of the relationship between socioeconomic status and reading achievement across countries.
- Underscored the link between achievement and school types and curriculum tracking within schools.
- Supported the notion that public and private schools tend to have the same effects for the same kinds of pupils but that private government-dependent schools are relatively more effective for pupils from lower socioeconomic levels.
- Stressed the need for intensive language and reading programs for foreign-born students to help boost achievement (Switzerland).

PROBLEMS WITH INTERNATIONAL ASSESSMENTS

Despite obvious advantages, a number of problems associated with international assessments merit consideration before countries decide to participate in one (see Kellaghan 1996).

First, an assessment procedure that will adequately measure the outcomes of a variety of curricula is difficult to design. Although curricula across the world have common elements, particularly at the primary-school level, considerable differences between countries also exist in what is taught, when it is taught, and what standards of achievement are expected.

South Africa's review of TIMSS items shows that only 18 percent of the science items matched the national curriculum of grade 7, while 50 percent matched the grade 8 curriculum (Howie and Hughes 2000). The greater the difference between the curricula and levels of achievement of countries participating in an international assessment, the more difficult it is to devise an assessment procedure that will suit all countries, and the more doubtful is the validity of any inferences that are made about comparative achievements.

We would expect an achievement test that is based on the content of a national curriculum to provide a more valid measure of curriculum mastery than would one that was designed to serve as a common

denominator of the curricula offered in 30 to 40 countries. For example, a national curriculum authority and the designers of an international assessment might assign quite different weights of importance to a skill such as drawing inferences from a text. A national assessment, as opposed to an international assessment, can also test curricular aspects that are unique to individual countries.

Devising a common assessment instrument is more difficult for some curriculum areas (for example, science and social studies) than for others (for example, reading). In the case of science, for example, achievement patterns have been found to be more heterogeneous than in mathematics. Furthermore, a greater number of factors are required to account for student performance differences in science than in mathematics. Thus, a science test that would be appropriate for a variety of education systems is difficult to envisage.

A second problem with international studies is that—although early studies had the ambitious aim of capitalizing on the variation that exists in education systems to assess the relative importance of a variety of school resources and instructional processes—this goal, in practice, turned out to be very difficult to achieve. Because the relative effect of variables depends on the context in which they are embedded, practices associated with high achievement in one country cannot be assumed to show a similar relationship in another. In fact, the strength of correlations between background factors and achievement has been found to vary from country to country (see, for example, OECD and UNESCO Institute for Statistics 2003; Wilkins, Zembylas, and Travers 2002). Particular difficulties exist when developing countries are involved in a study designed for industrial countries because socioeconomic factors in such countries can differ very much from those that prevail in industrial countries and can include poverty, nutritional and health factors, and poor educational infrastructure and resourcing.

Third, the populations and samples of students participating in international assessments may not be strictly comparable. For example, differences in performance might arise because countries differ in the extent to which categories of students are removed from mainstream classes and so may be excluded from an assessment (for example, students in special programs or students in schools in which the language of instruction differs from the language of the assessment).

The problem is most obvious where (a) age of enrolling in schools, (b) retention, and (c) dropout rates differ from one country to another and is particularly relevant in studies in which industrial and developing countries participate. In some developing countries, large proportions of students have dropped out well before the end of the period of compulsory schooling. Whereas primary school net enrollment ratios for Western Europe and North America are almost 100 percent, the ratios for countries in Sub-Saharan Africa are, on average, less than 60 percent (UNESCO 2002). Patterns of early dropout can differ from country to country. In Latin American and Arab countries, boys are more likely than girls not to complete grade 5; the reverse is true in some African countries (for example, Guinea and Mozambique). Sampling problems for TIMSS appeared in the Republic of Yemen, where several schools did not have grade 4 classes and where one school for nomadic children could not be located.

Similar comparability problems can arise in a national assessment. For example, the differential performance of students in states in India has been attributed to differential survival rates (see A.1 in appendix A).

Fourth, because variation in test score performance is an important factor if one is (a) to describe adequately the achievements of students in the education system and (b) to determine correlates of achievement, carefully designed national tests must ensure a relatively wide distribution of test scores. However, many items in international assessments have been too difficult for students from less industrial countries, resulting in restricted test score variance. This result is reflected in the data presented in table 6.2, which are based on a selection of countries that participated in TIMSS 2003.

The data show the percentage of grade 8 students who reached levels or benchmarks of performance when compared to all students who took the test. Seven percent of all those who took the mathematics test achieved the "advanced" international benchmark, 23 percent the "high" benchmark, one-half the "intermediate" benchmark, and roughly three-quarters the "low" benchmark. In sharp contrast, 2 percent of Ghanaian students achieved the "intermediate" benchmark and 9 percent achieved the "low" benchmark. Zero percent achieved the "advanced" and "high" international benchmarks.