NATIONAL AND INTERNATIONAL ASSESSMENTS OF STUDENT ACHIEVEMENT

Vincent Greaney
Thomas Kellagan

Abstract: This introductory book describes 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. The purposes and main features of national assessments are described in chapter 2. 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), are described.

Keywords: national and international assessment, analysis, students, achievement

Contents

1. Introduction

In this introductory volume we describe the main features of national and international assessments, both of which have become extremely popular in assessing the quality of education in the 1990s and 2000s. This increase in popularity reflects two important developments in the use of assessment data. First, it represents a shift from the traditional use of achievement tests to assess

individual students to their use to obtain information about the achievements of the system of education as a whole (or a clearly defined part of it). And secondly, it reflects a shift in emphasis in assessing the quality of education from a concern with inputs (e.g., student participation rates, physical facilities, curriculum materials, teacher training) to a concern with outcomes, that is the knowledge and skills that students have acquired as a result of their exposure to schooling (Kellaghan & Greaney, 2001).

National assessments are used to describe student learning achievement at a particular point in time and also to monitor changes in achievement over time. As part of its management function, a ministry of education should be able to describe national levels of learning achievement especially in key subject areas. It should also be able to use this information to compare achievement levels of key sub-groups, such as boys and girls, urban and rural students, and public and private school students. By having good quality achievement data at different points in time a ministry can also support or refute claims (and counter-claims) about rising or falling standards of student achievement.

The present series of booklets is designed to introduce readers to the complex technology that has grown up around the administration of national and international assessments. In this introductory volume, the purposes and main features of national assessments are described. Reasons for carrying out a national assessment are considered, as are the main decisions that have to be made in the design and planning of an assessment. Common mistakes in the design, implementation, analysis, and reporting of a national assessment are identified.

In the final chapter, international assessments of student achievement, which share many procedural features with national assessments (sampling, administration, the kind of background data collected, methods of analysis), are described. The main point of difference between the two is at once a strength and a weakness of an international assessment. It is a strength in that it provides data from a number of countries, allowing each individual country to compare its results with the results achieved by students in other countries; it is a weakness in that the requirement that test instruments be acceptable in all participating countries means that the instruments may not accurately reflect the achievements of students in individual countries.

A further feature of international assessments is the fact that participating countries usually carry out internal analyses based on data collected within a country. Thus, the data collected for the international study are used for what is in effect a national assessment. However, the practice is not without its problems, and data that are collected in this way may be less appropriate 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 are regional studies in which a number of countries in a region that may share many socioeconomic and cultural features collaborate in a study.

A further variation is a subnational assessment in which an assessment in confined to a province or state within a country. Subnational assessments have been carried out in a number of large countries (e.g., Argentina, Brazil, United States of America) to meet local or regional information needs. These differ from national assessments in which all participants use the same instruments and procedures, but which allow (if numbers are large enough) disaggregation of data for regions within the country in reporting results.

In a series of appendices to the present volume, descriptions of the main features of national assessments in eight countries are provided, followed by descriptions of three international studies and three regional studies.

Details in the design and implementation of a national assessment are considered in subsequent volumes in this series. The issues addressed are:

instrument development;
sampling, data preparation, and management;
data analysis;
logistics;
reporting and using national assessment results.

2. National assessments of student achievement

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 population of students, and may focus on a particular sector in the system (such as fifth grade or 13-year old students). Teachers and others (e.g., parents, principals) 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 such factors as levels of teacher training, attitudes towards curriculum areas, 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 /literacy and mathematics/ numeracy. Some assess students' achievements in science, a second language, art, music, or social studies. In practically all national assessments systems, students at the primary-school level are assessed. In many systems, national assessments are also carried out at secondary school, usually during the period of compulsory education.

National assessment systems also differ from country to country. They differ in the frequency with which assessments are carried out. In some, an assessment is carried out every year, though the curriculum area that is assessed may vary from year to year. In other systems, assessments are less frequent. A variety of agencies have been employed to carry out a national assessment: the Ministry of Education, a national research center, a consortium of educational bodies, a university, and an examination board. Participation by a school may be voluntary, or may be mandated. When voluntary, non-participation of some schools may bias the results and lead to an inaccurate reflection of achievement levels in the education system.

While most industrialized countries have had systems of national assessment for some time, it was only in the 1990s that the capacity to administer assessments became available in developing countries as a result of a shift in emphasis from educational inputs to outcomes in the assessment of quality following the 1990 Jomtien Declaration, World Declaration on Education for All. 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" (p. 5). More recently, the Dakar Framework for Action (UNESCO, 2000), produced at the end of the ten-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" (7, iv).

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. Ironically, the expectation that EFA and regular monitoring of achievement levels would result in an improvement in learning standards has not materialized. This may be because, while EFA led to rapid increases in numbers attending school, larger numbers may not have been matched by increased resources (especially trained teachers).

All 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, the aims of the curriculum, or preparation for life)?

– Is there evidence of particular strengths and weaknesses in students' knowledge and skills?

– Do particular subgroups in the population perform poorly? Are there, for

example, disparities between the achievements of boys and girls, of students in urban and rural locations, of students from different language or ethnic groups, of 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 (e.g., school resources, teacher preparation and competence, type of school) or with students' home and community circumstances?

– Are government standards being met in the provision of resources (e.g., textbooks, desks, 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. To answer the question, it will be necessary to carry out assessments that yield comparable data at different points in time (Kellaghan & Greaney, 2001, 2004).

Most of these questions were addressed in the design and implementation of Ethiopia's national assessment (Box 2.1).

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 quality education.

Source: Ethiopia. National Organization for Examinations (2005).

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

Box 2.2
Examples of Questions from 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 school buildings?
Questions related to standards of educational provision:
Were Ministry standards met regarding
– class size?
– classroom furniture?
– qualifications of staff?
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 variance in reading and mathematics achievement?
What were the major variables that differentiated between the most and least effective schools?

World Bank (2004). Vietnam: Reading and mathematics assessment study. Washington: Author.

What are the main elements in a national assessment?

While national assessments can vary in how they are implemented, they tend to have a number of common elements (Box 2.3) (Kellaghan & Greaney, 2001, 2004).

Box 2.3
Main Elements of a National Assessment

– The Ministry of Education (MOE) appoints an implementing agency either within the Ministry or an independent external body (e.g., a university department or a research organization) and provides funding.
– Policy needs to be addressed in the assessment are determined by the Ministry, sometimes in consultation with key educational stakeholders (e.g., teachers' representatives, curriculum specialists, business people, parents).
– The MOE, or a steering committee nominated by it, identifies the population to be assessed (e.g., fourth grade students).
– The area of achievement to be assessed is determined (e.g., literacy, numeracy).
– The implementing agency prepares achievement tests and supporting questionnaires and administration manuals.
– The tests and supporting documents are pilot-tested, and subsequently reviewed to determine curricular and technical adequacy.

– The implementing agency selects the targeted sample (or population) of schools/ students, arranges for printing of materials, and establishes communication with selected schools.
– Test administrators (e.g., classroom teachers, school inspectors, or graduate university students) are trained by the implementing agency.
– Survey instruments (tests and questionnaires) are administered in schools on a specified date.
– Survey instruments are collected, returned to the implementing agency, cleaned, and prepared for analysis.
– Data analysis is carried out.
– Draft reports are prepared and reviewed.
– The final report(s) is prepared and disseminated.

It is clear from the list of elements in Box 2.3 that a good deal of thought and preparation is required before students respond to assessment tasks. A body with responsibility for collecting data has to be appointed, decisions have to be made about the policy issues to be addressed, and tests and questionnaires have to be designed and tried out. In preparation for the actual testing, samples (or populations) of schools and of students have to be identified, schools have to be contacted, and test administrators selected and trained. In some countries (e.g., Vietnam and some African countries), teachers respond to the assessment tasks taken by their students. 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.

It is important that the student achievements that are assessed are considered to be important outcomes of schooling; that the method of sampling ensures that data that are collected adequately represent the achievements of the education system as a whole (or of a clearly identified part of it); and that analyses identify and describe the main features of the data that have been collected, including relationships between significant variables. All these activities require considerable resources and political support.

It should be recognized that low-income countries encounter problems over and above those encountered by other countries in attempting to carry out a national assessment. To begin, education budgets may be meager; some countries devote less than 1% of GDP to education (e.g., Azerbaijan., Cambodia, Central African Republic, Congo, Georgia, Haiti, Guinea, Guinea Bissau, Sudan) compared to over 5% in most middle-income countries. Competing demands within the education sector for activities such as school construction, teacher training, and provision of educational materials can result in non-availability 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 problems due to inadequate roads, mail and telephone services. Finally, the very high between-school variation in student achievement typically found in low-income countries requires a larger sample than is required in more developed countries.

How does a national assessment differ from public examinations?

It should not be assumed that public examinations can provide the information that a national assessment can and that if a country has a public examination system, there will not be a need for a national assessment system.

A number of features of public examinations mean that they cannot provide the kind of information that a national assessment seeks to provide. First, since public examinations play a major role in the selection of students (for the next highest level in the education system, and sometimes for jobs), they may not provide adequate coverage of the curriculum. Second, examinations, and the characteristic of students who take them, change from year to year, making comparisons over time very difficult. Third, the fact that "high stakes" are attached to performance (i.e., how students do on an examination has important consequences for them, and perhaps for teachers) has implications for the validity of the examinations. 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, 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.

Box 2. 4 summarizes the major differences between national assessments and public examinations.

Box 2. 4
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 (e.g. 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, etc.	Low importance	Great importance
Coverage of curriculum	Generally confined to one or two subjects	Cover main subject areas
Effect on teaching	Very little direct effect	Major: teachers tend 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 based on a predetermined marking scheme
Impact on level of student attainment	Unlikely to have impact	Poor results or the prospect of failure can lead to early dropout.
Usefulness for monitoring trends in achievement levels over time	Appropriate if tests are designed with this in mind	Not appropriate as examination questions and candidate populations vary from year to year.

How does a national assessment differ from classroom assessment?

Classroom assessment is an integral part of the teaching-learning process. In addition to ongoing teacher observation, it also includes classroom questioning, quizzes, and marking of homework. It occurs during learning and is designed to assess students' level of knowledge, to diagnose problems, and make decisions about the next instructional step. If well done, it can provide regular evidence of the learning of individual students and can cover a range of areas apart from memory work. A national assessment, on the other hand, is not primarily concerned with the achievements of individual students, or with any particular problems they may be experiencing. Data from individual students are aggregated to provide a picture of achievement in the education system in general. Furthermore, a

national assessment usually provides a measure of student achievement in only one, two, or three core curriculum areas using a limited range of question formats as infrequently as once every three or four years.

3. Why carry out a national assessment?

There are a variety of reasons why a decision might be made to carry out a national assessment.

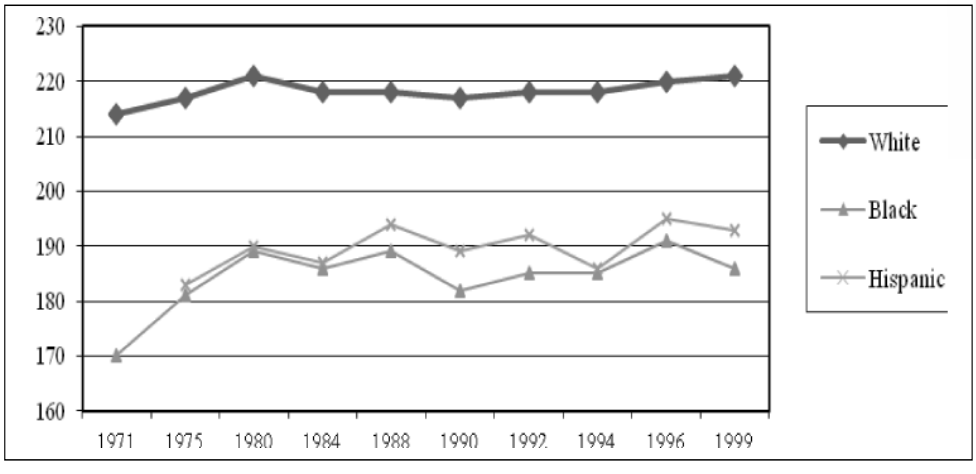
First, an assessment may primarily play a symbolic role to legitimate state action by embracing internationally accepted models of modernity and imbueing the policy-making process with the guise of scientific rationality (Benveniste, 2000; Benveniste, 2003; Kellaghan, 2003). When this is the motivation, the act of assessment is of greater significance than its outcomes. When a national assessment is carried out simply to meet the requirement of a donor agency, it may have little more than symbolic value and its findings may not be seriously considered in management of the education system or in policy making.

Secondly, a national assessment may reflect the efforts of a government to "modernize" its educational system by introducing a business management (corporatist) approach. This approach draws on concepts used in the world of business, such as strategic planning, a focus on deliverables and results, and 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 considered to be of interest to policy makers, politicians, and the broader educational community.

An important aspect of this approach is simply to provide information on the operation of the education system. Many governments lack basic information on key aspects of the system, including achievement levels and basic inputs to the system. National assessments can provide such information, a key prerequisite for sound policy making. For example, Vietnam's national assessment helped establish that many classrooms lacked basic resources. In a similar vein, Zanzibar's assessment reported that 45% of pupils lacked a place to sit. Bhutan's national assessment noted that some students had to spend a number of hours each day travelling to and from school. Namibia's assessment showed that many teachers had limited mastery of basic skills in English and mathematics.

An extension of describing conditions is to determine whether standards improve, disimprove, or remain static over time. A series of studies carried out in Africa established that between 1995/96 and 2000/01 there was a significant decline in reading literacy scores in Malawi, Namibia, and Zambia. In the United States, the National Assessment of Educational Progress (NAEP) monitored levels of reading achievement over almost three decades. It found that while nine-year old black and Hispanic children initially reduced the achievement gap with whites, the test score

differential remained fairly constant thereafter (Figure 3.1). In the U.S. also, NAEP helped identify the changing levels of reading achievement in states (Figure 3.2).



Source: Winograd & Thorstensen, 2004.

Figure 3.1
The Achievement Gap in the United States
Age 9 Students NAEP Reading Assessment 1971-1999

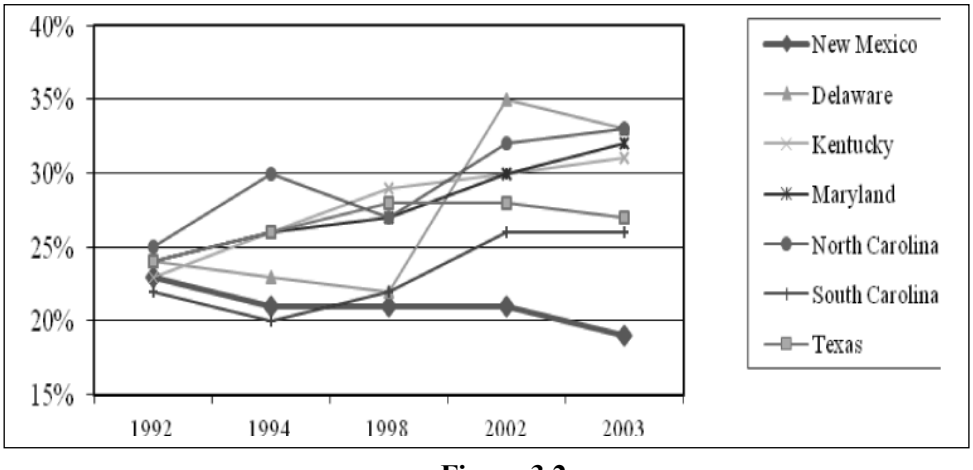


Figure 3.2
4th Grade Students At or Above Proficient in Reading
NAEP 1992–2003

The use that can be made of national assessment data depends on whether data were collected in a sample of schools or in a census in which information is available about all (or most schools). In both cases, results can be used to provide direction to policy makers interested in enhancing educational quality. They can help governments identify the strength of the association between the quality of student learning and various factors over which they have some control (e.g., availability of textbooks, class size, and number of years of teacher pre-service training). Analysis of findings can lead to decisions affecting the provision of resources in the education system in general (e.g., for the reform of curricula and textbooks, for teacher development) or in categories of school with particular characteristics (e.g., schools in rural areas, schools serving students in socioeconomically disadvantaged areas). Many examples can be found of the use of the findings of national and international assessments for these purposes. For example, they have been used to give direction to the providers of inservice education for teachers in Africa; they have prompted curriculum reform in Eastern Europe, have helped divert financial resources to poorer schools in Chile, and have promoted teacher professionalism in Uruguay.

The results of a national assessment may also be used to change practice in the classroom. However, getting information to teachers and effecting changes in their behaviour that will substantially raise the achievements of students is not an easy task. The pressure on schools and classrooms to change is greatest when the results of a national assessment based on a census are available and high stakes are attached to performance. In this case, no specific action may be taken beyond the publication of information about performance (e.g., in league tables). In some cases, sanctions may be attached to performance. These can either take the form of inducements for improved performance (e.g., schools and/or teachers receive money if students achieve a specific target) or 'punishments' for poor performance (e.g., non-promotion of students, dismissal of teachers).

When a national assessment obtains information that can be used to make stakeholders accountable for student learning, usually it is assumed that teachers and schools bear the major responsibility. The focus is seldom on the accountability of those who are responsible for determining educational policies or for providing the funding to implement policies and programs.

4. Decisions in a national assessment

In this section, we consider the series of decisions that are involved in planning a national assessment.

1. 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 committee which will actually carry out the assessment. Such a 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 a NSC will vary from country to country depending on the power structure within the education system. In addition to the Ministry of Education, NSCs might include representatives of major ethnic, religious, and linguistic groups as well as those whose members will be expected to act on the results, such as teacher trainers, teachers, school inspectors, and curriculum personnel. Addressing the information needs of these various stakeholders should help to 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 or should not be required to address routine implementation tasks related to the national assessment. In some cases, it may provide direction at the initial stage by identifying the purpose and rationale of the assessment, determining the subjects and grade level (s) to be assessed, or select the agency or agencies to conduct the assessment, though these may also be decided before the committee is established. The NSC is likely to be most active at the start of the assessment exercise while the implementing agency will be responsible for most of the detailed work such as instrument development, sampling, analysis, and reporting. The agency, however, should provide the NSC with draft copies of tests and questionnaires and descriptions of proposed procedures so that committee members can provide guidance and 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.

2. 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 non-national technical teams. One would expect a variety of factors to influence such a decision, including levels of national technical capacity, as well as administrative and political circumstances. For example, while autonomous bodies may have a higher degree of functional independence and technical legitimacy, they may be disconnected from ministry information needs, and so have little impact on policy. Some potential advantages and disadvantages of different categories of implementation agencies

which merit consideration in deciding who should carry out an assessment are listed in Box 4.1.

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 technical capacity. At a later stage, provinces were presented with a set of sample questions for their endorsement and the National Direction of Evaluation (NDE) constructed the final assessment instruments from the pool of pre-approved test items. More recently, test items have been designed independently by university personnel and approved by the national Federal Council. The NDE remains responsible for the design of achievement tests, the analyses of results, and the general co-ordination of annual assessment activities.

Box 4.1
Options for Implementing a National Assessment

Designated agency	Advantages	Disadvantages
1. Drawn from staff of Ministry of Education	Likely to be trusted by Ministry. Ready access to key personnel, materials, and data (e.g., school population data). Funds for staff time may not have to be secured.	Findings might be subject to political manipulation including suppression. May be viewed skeptically by other stakeholders. Staff may be required to undertake many other tasks. Technical capacity may be lacking.
2. Drawn from staff of Public Examination Unit	Usually credible. Experience in running secure assessments. Funds for staff time may not have to be secured. Some skills (e.g., test development) can be transferred to enhance the Examination Unit. More likely to be sustainable than some other models.	Staff may be required to undertake many other tasks. Technical capacity may be weak. May lack ready access to data. Public examination experience may result in test items that are too difficult.
3. Drawn from research-university sector	Findings may be more credible with stakeholders. Greater likelihood of having 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.

4. Recruited as foreign technical assistance (TA)	More likely to recruit a technically competent team. Nature of funding 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.
5. Made up of a national team supported with some international technical assistance (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.
6. Ministry team supported with National TA	Can ensure Ministry support while obtaining national TA. Less expensive than international TA.	National TA may lack the technical capacity. Other potential disadvantages outlined in No 1 above may apply.

Less expensive than international TA. National TA may lack the technical capacity.

Other potential disadvantages outlined in No 1 above may apply.

It is worth reflecting on the wide variety of skills that is required to carry out a national assessment in deciding who should be given responsibility for the task. Furthermore, a national assessment is fundamentally a team effort. The team should be flexible, willing to work under pressure and in a collaborative manner, and be prepared to learn new assessment and technological approaches. The team leader, sometimes termed the National Research Coordinator (NRC), should have strong managerial skills. He/she will be required to organize staff, coordinate and schedule activities, support training, and arrange and monitor finance. Given the need to report to a national steering committee, liaise with national, regional and in some instances district-level government bodies, and representatives of stakeholders such as teachers and religious bodies, the coordinator should be politically astute.

The team should have high-level implementation or operational skills. Tasks to be completed include the development of training materials; the organization of workshops for item writers and test administrators; arranging for printing and distribution of tests, questionnaires, and manuals; contacting schools; 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 pre-testing 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 preparing data files for others (e.g., academics and post-graduate students) to carry out secondary analyses.

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 educational officials to discuss the importance of the results and their implications for teaching and learning.

Most members of the team may be part-time and employed "as needed". This category could include item writers, especially practicing teachers with a good knowledge of the curriculum. It might also include experts in areas such as 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 with drafting catchy press releases.

3. 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 the national assessment tests and questionnaires. Practice varies. For example, Colombia has used graduate students while Zambia has involved school inspectors and Ministry officials in test and questionnaire administration. In Argentina, each province selects its own independent proctors. 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.

4. At what level of schooling will students be assessed?

Policy makers want information about the knowledge and skills of students at selected points in their educational careers. A decision that has to be made is whether populations are defined on the basis of age or grade, or indeed a combination of age and grade. In countries in which 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 towards the end of the last year of primary school (5th or 6th grade in many countries). The ministry could also request a national assessment in grades 3 or 4 if it needs data on how students are performing midway through the basic education cycle. This information could then be used to introduce remedial measures (e.g., 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 in grades 4, 8, and 12; in Colombia, achievement is assessed at grades 3 and 5; in Uruguay, at grades 1, 2, and 6; in Sri Lanka, at grades 4, 8, and 10. In Anglophone Africa, a regional consortium of educational systems, the Southern Africa Consortium for Monitoring Educational Quality (SACMEQ) assessed grade 6 students. Countries in the Francophone Africa consortium, Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC), assess 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 as any lower level would have required translation of tests into 270 local languages. More senior grades were not considered suitable as 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 this level might not be able to follow instructions, or cope with the cognitive tasks of the assessment or with the challenge of completing multiple-choice tests. A Jamaican study noted that a sizeable number of grade 1 students were unable to recognize the letters of the alphabet (Lockheed & Harris, 2005).

5. Will a whole population or a sample be assessed?

The purpose of an assessment is key to determining whether or not to test a sample or the entire population of targeted students. Cost also is a factor. For obvious reasons, samples are less expensive than population-based approaches. Most national and all regional and all international studies use sample-based approaches in determining national achievement levels. Some national assessments use both census and sample-based approaches, while most subnational assessments collect census data.

The decision to involve an entire population in a national assessment may reflect an intention to foster school, teacher, or even student accountability. It facilitates the use of sanctions (incentives, penalties), the provision of feedback on performance to individual schools, and the publication of league tables, as well as the identification of schools with the greatest need for assistance (e.g., in Chile and Mexico). The sample-based approach, on the other hand, will only permit the detection of problems at the system level. It will not identify specific schools in need of support, though it can identify types or categories of school (e.g., small rural schools) that require attention. It can also identify problems relating to gender or ethnic equity.

In some countries, it may be difficult to define the population of schools or subpopulations to be assessed because of lack of up-to-date information on functioning

schools. Information on schools may not be accurate. Functioning schools may not appear on all official lists, while non-existent (ghost schools) may be listed. Detailed information on pupil enrolment in individual grades may also be lacking or inaccurate.

6. What will be assessed?

All national assessments measure cognitive outcomes of instruction or scholastic skills (language/literacy and mathematics/numeracy), a reflection of the importance of these outcomes for basic education. In some countries, science and social studies are included in an assessment. Whatever the domain of the assessment, it is important that an appropriate framework be provided, in the first instance for the construction of assessment instruments and afterwards for the interpretation of results. The framework may be available in a curriculum if, for example, the curriculum provides expectations for learning which are clearly prioritized and operationalized. In most cases, however, such a framework will not be available, and those charged with the national assessment will have to construct it. In this task, close cooperation will be required between the assessment agency, those responsible for curricula, and other stakeholders.

An alternative to basing an assessment instrument on curriculum-embedded expectations or even 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. Thus, for example, the Programme for International Student Assessment (PISA) 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, p. 24). This approach fitted well in an international study, since the alternative of devising an assessment instrument that would be equally appropriate to a variety of curricula is obviously problematic. However, it might also be used in a national assessment.

A few national assessments have collected information on affective outcomes (e.g., student attitudes to school, student self-esteem). However, measures of these outcomes tend to be unreliable, and analyses based on them have proved difficult to interpret.

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

Identification of contextual factors related to student achievement can help in the identification of manipulable variables (e.g., time allocated to curriculum areas, the availability of textbooks) that may affect student learning and in the determination of policy relating to the allocation of financial resources.

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

7. How will achievement be assessed?

It is necessary to devise an instrument or instruments that will provide the information that the national assessment was designed to obtain. Since the purposes and proposed uses of national assessments vary, so too will the instruments used in the assessments.

A major distinction is often drawn between norm-referenced and criterion-referenced tests. If an assessment is designed primarily to compare the achievements of different groups, it might follow the procedures for the construction of norm-referenced tests. It should be noted that the use of such tests does not preclude the possibility of examining student achievements in some detail (e.g., categorized by achievement domain). More explicit efforts may be made in instrument construction, however, to identify the extent to which students have acquired expected knowledge and skills (with perhaps judgments about the proportion of students whose achievements can be described as 'minimum' or 'satisfactory'), in which case a more criterion-referenced approach will be adopted.

In practice, instrument development often involves aspects of both a norm-referenced and a criterion-referenced approach. 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. However, if items do not meet certain criteria when tried out, in particular criteria relating to difficulty level, they may not be included in the final assessment instrument. Whatever approach is adopted, care should be taken to ensure that important curriculum objectives are reflected in an assessment, even if all or no students in the tryout provide evidence of achieving them.

Most national and international assessments rely to a considerable extent on the multiple-choice format in their instruments. However, these items will often be supplemented by open-ended items that require the student to write a work, phrase, or sentence.

In several national (e.g., US NAEP, Ireland) and international assessments (e.g., TIMSS, PISA), students respond to only a fraction of the total number of items used

in an assessment. This approach increases overall test coverage of the curriculum, without placing too great a burden on individual students. It also allows for the use of extended passages (e.g., a short story or a lengthy newspaper article) in the assessment of reading comprehension. In other assessments, all students respond to the same set of items. While there are advantages associated with having students respond to only a fraction of items, there are also disadvantages, particularly for countries beginning a national assessment program. Administration (e.g., printing, distribution) is more complex, as is scoring and scaling of scores, while analyses involving individual student or school data can be problematic (see Sofroniou & Kellaghan, 2004).

8. How frequently will assessments be carried out?

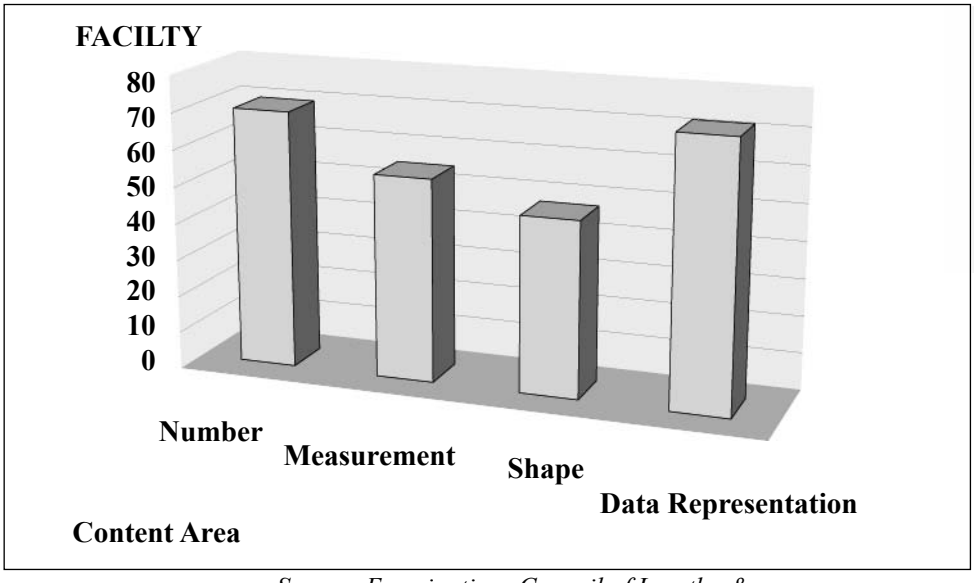
The frequency with which a national assessment is carried out varies from country to country, from every year to every ten years. There may be a temptation to assess achievement in the same curriculum areas and in the same population every year. This is unnecessary, as well as very expensive, if the aim is to monitor national standards. In the US, 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 (2001-2006). In Japan, achievement in core curriculum areas is assessed every ten 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, since 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.

However, if the purpose of an assessment is only to provide information on the performance of the system as a whole, an assessment of a sample of students in a particular curriculum area every three to five years would seem adequate. As educational systems do not change rapidly, more frequent assessments would be unlikely to register change. Over-frequent assessments would more than likely limit the impact of the results as well as incurring unnecessary costs.

9. How should student achievement be reported?

A variety of procedures have been used to describe student achievements in national assessments (see Volume 5 of this series). These range from the simple reporting of the percentage of students answering individual items correctly to the percentage of students performing at varying levels of 'performance standards' (e.g., 'basic', 'proficient', 'advanced'). Between these two extremes, performance may be reported in terms of curriculum domains or content areas (see Figure 4.1) or of 'mastery' of curriculum objectives.



Source: Examinations Council of Lesotho &
National Curriculum Development Centre (2006)

Figure 4.1
Lesotho: Grade 6 Math Performance by Content Area

10. What kinds of statistical analysis 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/rural location, ethnic or language group membership, and type of institution attended (public/private). Analyses may also throw light on the school and background factors that contribute to achievement, and so merit consideration in educational policy-making and decision-making.

The limitations of these 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. It is difficult to disentangle the effects of community, home, and school factors on students' learning. This has not deterred some investigations from assuming that data collected in national and international assessments can be interpreted causally. For example, in a study using TIMSS data in South Africa entitled 'System-level evaluation: Language and other factors affecting mathematics achievement', the question posed for analysis was: 'What factors in school level, class level and student level influence pupils' performance in mathematics?' (Howie, 2005, p. 177). Use of the words 'affecting' and 'influence' imply a cause and effect relationship that is not warranted.

The fact that students in larger classes achieve at a higher level than students in smaller classes is sometimes interpreted to mean that a reduction in class size would do nothing to improve achievement. However, again the conclusion may not be warranted if, for example, closer analysis revealed that smaller classes are found in rural areas, and have poorer resources including less qualified teachers, while larger classes, located in urban schools have better resources, including more highly qualified teachers. They probably also include more students from relatively advantaged home backgrounds.

11. How should the results of a national assessment be communicated and used?

If the reports of a national assessment are to have an impact on national educational policy, they should be produced as soon as possible after the completion of data analysis. In the past, technical reports which featured a considerable amount of data and technical terminology tended to be the sole form of reporting. It is now increasingly recognized that other forms of reports are required. These include short summary reports for busy policy makers which focus on the main findings, press releases, special reports for radio and television, and separate reports for groups such as curriculum developers and teacher trainers. In some countries (e.g. Sri Lanka), separate reports are prepared for each province. The information needs of readers should determine the contents of additional reports. Volume 5 in this series has an extensive section on report writing.

National assessment results have been used to set benchmarks for monitoring learning achievement levels (e.g. Lesotho), curriculum reform, providing baseline data on the amount and quality of educational materials in schools (e.g., Vietnam), for identifying correlates of achievement, and for diagnosing aspects of the curriculum which 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 behavioural areas that subsequently helped direct a large-scale teacher in-service program.

5. Common mistakes in the design, implementation, analysis, and reporting of a national assessment

Serious mistakes can undermine the degree of confidence placed in results and can render them of limited value or worthless for policy making. In this section, we identify some common mistakes that have been made in national assessments that have been carried out to date at the planning, implementation, analysis, report writing, and use of results stages. National assessment teams should study these and take steps to avoid making similar mistakes in their own national studies.

Planning Mistakes

– Failing to make adequate financial provision for key aspects of a national assessment, including report writing and dissemination.

– Failing to set up a national steering committee, and use it as a source of information and guidance during the course of the national assessment.

– Lack of government ownership and commitment to the process of national assessment reflected in the absence of a national steering committee or in simultaneous competing ongoing national assessments (often supported by external donors).

– Failing to involve key stakeholders (e.g. teachers' representatives, teacher trainers) in the planning of the national assessment.

– Omitting a key sub-group from the population assessed (e.g., private schools, students in remote areas).

– Setting unrealistic test score targets (e.g., 25% increase in scores over a four-year period).

– Allowing inadequate time for test development.

Implementation Mistakes

– Assigning test development tasks to people who are unfamiliar with the likely levels of student performance (e.g., academics), resulting in tests that are too difficult.

– Inadequate representation of the curriculum in tests.

– Failing to pilot test items.

– Using an inadequate number of test items in the final version of the test.

– Failing to design items that compromise between what students should know and what they actually know, resulting in items that are too difficult.

– Failing to give a clear definition of the construct being assessed (e.g., 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 prior to taking the test.

– Failing to give adequate notification to printers of tests, questionnaires, and manuals.

– Paying insufficient attention to proof reading tests, questionnaires, and administrative manuals.

– Using inadequate/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 (e.g., 5%).

– Allowing local administrators and school officials too much flexibility in selecting schools and students to participate in the national assessment. Giving inadequate training to test and questionnaire administrators.

– Allowing outside intervention (e.g., principal sitting in the classroom) during test administration.

– Allowing students sit close to each other during the assessment (encourages copying).

– Failing to establish a tradition of working outside normal work hours to complete key tasks on time.

Analysis Mistakes

– Using inappropriate statistical analyses, including failing to weight sample data in analysis.

– Basing results on small numbers (e.g., a minority of sampled teachers might have responded to a particular question).

– Contrasting student performance in different subject areas and claiming that students are doing better in one subject area based on mean score differences.

– Failing to emphasize the arbitrary nature of selected test score cut-off points (e.g., mastery/non mastery, pass/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 based on achievement test results without taking into account key contextual factors which contribute to the ranking. Different rankings emerge when school performances are compared using unadjusted performance scores, scores adjusted for contextual factors (e.g., the percentage of students entitled to free school meals), and scores adjusted for earlier achievement.

– Inferring causation where it might not be justified, e.g., attributing differences in learning achievement to one variable (e.g., private school administration, class size).

– Comparing test results over two time periods even though non-equivalent test items were used.

– Comparing test results over two time periods without reporting the extent to which important background conditions (e.g., curriculum , enrollment, household income, or level of civil strife) might have changed in the interim. While most education-related variables tend not to change rapidly over a short time period (e.g., 3- 4 years), some countries have introduced policies which have resulted in major changes in enrollment. Uganda, prompted by the EFA agenda, greatly increased the number of students enrolling in schools. Malawi's policy of dropping school fees also led to a rapid enrollment increase.

– Limiting analysis in the main to a listing of mean scores of geographical or administrative regions.

Report-Writing Mistakes

– 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.

– Failure 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 over-extensive review of literature in the assessment report.

– Failing to publicize the key relevant messages of the report to separate stakeholder audiences.

Mistakes with Results

– Ignoring the results when it comes to policy making.

– Failure of key stakeholders (e.g. teacher trainers, curriculum personnel) to consider the implications of the national assessment findings.

– Failure of the national assessment team to reflect on lessons learned and to take note of these in follow-up assessments.

Box 5.1 contains a summary of points which might be used as a check list to help improve the overall quality of national assessment design, implementation, analysis, report writing, and follow-on activities. Not all of these are likely to be appropriate for every national assessment, as education systems tend to operate under different political, financial, and administrative constraints.

Box 5.1 Checklist of Suggestions for Improving the Quality and Impact of National Assessments
Activity
1. Involve senior policy makers from the outset to ensure political support and to help frame the assessment design.
2. Address the information needs of policy-makers when selecting aspects of the curriculum, grade levels, and population subgroups (e.g., by region, by gender) to be assessed.
3. Obtain teacher support by involving their representatives in assessment-related policy decisions.
4. Be aware that attaching high stakes to students' performance may lead to teacher opposition, and a narrowing of the effective curriculum to subjects or aspects of subjects that are assessed.
5. Define precisely the constructs and skills to be assessed.
6. Secure the services of a person/unit with sampling expertise.

7. Specify the defined target population (the population from which a sample will actually be drawn) (the sampling frame), and the excluded population (e.g., elements of the population that are too difficult to reach or that would not be able to respond to the instrument).
8. 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.
9. Select members of the sample from the sampling frame according to known probabilities of selection.
10. Entrust test development to personnel who are familiar both with curriculum standards and learning levels of students (especially practicing teachers).
11. Use assessment instruments that adequately assess the knowledge and skills about which information is required, and which will provide information on sub-domains of knowledge or skills (e.g., problem solving) rather than just an overall score.
12. Develop clear and unambiguous test and questionnaire items and present them in a clear and attractive manner.
13. Pilot test and revise items, questionnaires, and manuals.
14. Proof-read all materials carefully.
15. Follow a standard procedure when administering tests and questionnaires. Prepare an administration manual.
16. Ensure that test administrators are thoroughly familiar with the contents of tests, questionnaires and manual, and with administrative procedures.
17. Prepare and implement a quality assurance mechanism to ensure that administration procedures are followed.
18. Secure competent statistical services.
19. Prepare a codebook with specific directions for preparing data for analysis.
20. Check/clean data to remove errors (e.g., relating to numbers, out-of range scores, and mismatches between data collected at different levels).
21. Calculate sampling errors taking into account complexities in the sample, such as stratification and clustering.
22. Weight data so that the contribution of the various sectors of the sample to aggregate achievement scores reflects their proportions in the target population.
23. Identify the percentage of students who met defined acceptable levels/standards.
24. Report results by gender and by region, if sample design permits.
25. Analyze assessment data to identify factors that might account for variation in student achievement levels or for research studies to help frame policy making
26. Analyse results by curriculum domain. Provide information on the sub-domains of a curriculum area (e.g., aspects of reading, mathematics).

27. Recognize that it is not usually possible to make direct comparisons between performances in different curriculum areas such as reading and mathematics.
28. Recognize that a variety of measurement, curricular, and social factors may account for student performance.
29. 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.
30. Provide adequate information in the report or in a technical manual to allow for replication of the assessment.
31. Use results to provide direction for pre-and in-service teacher education courses and for curriculum authorities.

6. International assessments of student achievement

An international assessment of student achievement is similar in many ways to a national assessment. Both exercises make use of similar procedures (in instrument construction, sampling, scoring, and analysis). They also may have similar purposes: to determine how well students are learning in the education system; to identify particular strengths and weaknesses in the knowledge and skills that students have acquired; to compare the achievements of subgroups in the population (e.g., defined in terms of gender or location); to determine the relationship between student achievement and a variety of a characteristics of the school learning environment and of homes and communities. Furthermore, both exercises may attempt to establish if student achievements change over time (Kellaghan & Greaney, 2004).

The main difference between an international and a national assessment is that the former is carried out in more than one country and 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. 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. Much of the interest in international assessments can also be attributed to the belief that human capital (in particular those aspects of it represented by mathematics and science achievements) plays an important role in economic growth. As a consequence, education policy around the world has increasingly focused on improving aggregate student achievement as a means to increase economic growth. There is some research evidence to support this approach, though it is not entirely consistent across countries or over time (Hanushek & Kimko, 2000; Ramirez, Luo, Schofer, & Meyer, 2006).

The strength of the belief in the importance of human capital, and that international studies adequately assess it, is reflected in the fact that since the

1960s, over 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 the IEA studies Trends in International Mathematics and Science (TIMSS) and Progress in International Reading Literacy Study (PIRLS) and the OECD Programme for International Student Assessment (PISA). Regional international assessments in reading and mathematics have been carried out in Southern and Eastern Africa (SACMEQ), Francophone Africa (PASEC), and Latin America (LLECE).

The results of international assessments (TIMSS and PISA) and regional assessments can and have been used to prepare separate national reports on country-level performance. International data bases 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 developed countries conduct their own national assessments as well as participating in international assessments. The United States has its own National Assessment of Educational Progress (NAEP) for grades 4, 8, and 12 and also participates in international assessments of achievement. Some countries participated in international assessments but did not conduct national assessments (e.g., the Russian Federation, Germany). India, on the other hand, has run large-scale national assessments for some decades but, with one exception, has not participated in a major international assessment. Many of the world's poorest countries do not participate in international assessments or carry out national assessments, though the situation has changed in recent years.

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 (e.g., students in two adjacent grades that contain the largest proportions of 9-year olds and 13-year olds in TIMSS; students in the grade levels containing most 9-year olds and most 14-year olds in the IEA Study of Reading Literacy; the upper of two adjacent grades with the most 9-year olds in PIRLS). In another study, separate age and grade samples were selected (the 1964 IEA First Mathematics Study). In yet another, students of a particular age were selected (15-year olds in PISA).

The number of countries participating in international studies has increased over the years. While typically less 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, PISA) are administered on a cyclic basis and are now described as 'trend' studies.

Participation by non-industrialized countries in international studies has generally been low. However, in line with the general increase in the number of countries that have taken part in international studies, the number has increased over the years. TIMSS attracted the largest number in 2003 (seven from Africa and nine from Asia and the Middle East). As was the case generally in international studies, non-industrialized countries have shown a greater interest in taking part in studies of mathematics and reading than in studies of other curriculum areas.

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. In several studies, comparisons of curricula have led to decisions to increase the focus on a curriculum area or to include or exclude curriculum material. In South Africa, for example, poor achievement results from TIMSS led to increased allocation of resources for science and mathematics (Reddy, 2005). The results of a national assessment may also identify curriculum areas (e.g., formal grammar, problem solving in mathematics, science) which differ from curricula in other countries, or have not kept up to date with developments in other countries. Following such an analysis, countries can decide to alter aspects of the curriculum or, having reflected on alternatives, to stick with those aspects of the curriculum that are considered appropriate for local conditions.

By comparing results from different countries, assessment results can also be used to help define what is achievable, how achievement is distributed, and relationships between average achievement and its distribution. For example, can high average achievement co-exist with narrow disparities in performance?

International studies are likely to attract the attention of the media and of a broad spectrum of stakeholders (politicians, policymakers, academics, teachers, the public). Differences between countries in levels of achievement will be obvious in the descriptive statistics provided in reports of the studies and, indeed, these are usually highlighted in 'league tables' in which countries are ranked in terms of their mean level of achievement. The comparative data provided in the studies will have more 'shock value' than the results of a national assessment. Poor results can encourage debate, which in turn may provide politicians and other policymakers 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.

Data on achievement provide only limited information. On the basis that international studies can capitalize on the variability that exists across education systems, broadening the range of conditions that can be studied beyond those operating in any one country, analyses of data collected in international studies routinely consider associations between achievement on the one hand and a wide range of contextual variables (system-wide, school-level, and student-level) on the other. The range of variables considered include 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, as well as a basis for the evaluation of policy and practices.

International studies also have the potential to bring to light concepts for understanding education that have been overlooked in a country (e.g., in definitions of literacy, in conceptualizing curricula in terms of intention, implementation, and achievement). They can also help identify and lead to questioning assumptions that may be taken for granted (e.g., the value of comprehensive vs. selective education; that smaller class sizes are associated with higher achievement; that grade repetition benefits students).

International studies may also have a role to play in monitoring trends in achievement over time. This, of course, can also be achieved in a national assessment.

Finally, studies may contribute to the development of local capacity in a variety of technical activities: sampling, defining achievements, developing tests, statistical analysis, and report writing. Furthermore, staffing requirements and costs (for example, for instrument development, data cleaning, and analysis) may be lower than in national assessments since costs are shared with other countries.

Problems with International Assessments

Despite these obvious advantages, a number of problems associated with international assessments merit consideration before making a decision whether or not to participate in one.

First, it is difficult to design an assessment procedure that will adequately measure the outcomes of a variety of curricula. Any assessment can only incorporate a partial representation of any aspect of educational achievement. Although there are common elements in curricula across the world, particularly at the primary school level, there are also considerable differences between countries in what is taught and in expected standards of achievement.

South Africa's review of TIMSS items showed that only 18% of the science items matched the national curriculum of grade 7, while 50% matched the grade

8 curriculum (Howie & 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 the validity of any inferences that are made about comparative achievements. Qualitative differences between curricula and in the structure of students' achievements will not be adequately reflected in quantitative comparative data.

One would expect an achievement test based on the content of a national curriculum to provide a more valid measure of curriculum mastery than one which was designed to serve as a common denominator of the curricula on offer 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 which are unique to individual countries.

It may be noted that assessment studies generally reflect the values of industrialized countries and so may not take sufficient account of the national goals of participating developing countries which in the case of South Africa, for example, include transformation goals of access, redress, equity, and quality (Reddy, 2005).

It would seem more difficult to devise a common assessment instrument for some curriculum areas (e.g., science, social studies) than for others (e.g., 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 is required to account for student performance differences in science than in mathematics. Thus, it is difficult to envisage a science test that would be appropriate for a variety of education systems.

A second problem with international studies is that while early studies had the ambitious aim of assessing the relative importance of a variety of school resources and instructional processes, in practice this turned out to be very difficult to do. A major problem is that, since most studies are cross-sectional, their findings cannot be interpreted as representing causal relationships. This means that data from the studies relating achievement to background variables can, at best, only provide clues about what variables might be worth considering for manipulation. Further, they cannot give any assurance on what the effects of any manipulation might be.

Here, it may be noted that school and out-of-school contextual and socioeconomic factors can be very different in developing countries from those that prevail in industrialized countries, and include poverty, nutritional and health factors, and poor educational infrastructure and resourcing. However, a study designed for industrialized countries may not accord these adequate recognition.

A third problem relates to difficulties in analysis and interpretation arising from the complexity of interdependencies that exist among variables. A related issue is how to present the results of analysis that may be extremely complex in a form that will be intelligible to non-specialists. This is also a concern in national assessments, but the fact that the relative effect of variables depends on the context in which they are embedded adds to the complexity in international studies. Thus, it cannot be assumed that practices associated with high achievement in one country will 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.

Fourth, analytic problems are also created by the fact that in international assessments (and increasingly in national assessments) each student takes only a fraction of a large number of assessment tasks. While this has the advantage that it is possible to extend the range of curriculum coverage in tests without increasing the burden on individual respondents, it has the disadvantage that the sample of achievement data obtained from individual students may be less than satisfactory when used in analyses which seek to describe relationships between the achievements of individual students and other factors (see Sofroniou & Kellaghan, 2004).

Fifth, a particular problem arises in both international and national assessments if it is necessary to translate the instruments of the assessment into one or more languages. If comparisons are to be made between performances assessed in different languages, it should be realized that the 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 South Africa, words such as gasoline (petrol) and flashlight (torch) had to be changed. Ghana replaced the word "snow" with "rain". Problems involving calendar time can pose a difficulty in Ethiopia where the Coptic calendar has a thirteen-month year. It is not always possible to ensure, however, that the way questions are phrased and the cultural appropriateness of content are equivalent in all language versions of an assessment task. For example, material which is context-appropriate for African students, covering hunting, the local market place, agricultural pursuits, and local games, might be relatively unfamiliar in middle and high income countries.

Sixth, the populations and samples of students participating in international assessments may not be strictly comparable. 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 (e.g., students in special programmes, students in schools in which the language of instruction differs from the language of the assessment). The problem is most obvious where retention and dropout rates

differ from one country to another, and is particularly relevant in studies in which industrialized 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. While primary school net enrolment ratios for Western Europe and North America are almost 100 per cent, the ratios for countries in Sub-Saharan Africa are, on average, less than 60 per cent (UNESCO, 2002). Patterns of early drop-out 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 (e.g. Guinea, Mozambique).

Seventh, variation in test score performance is an important factor in determining correlates of learning achievement. While carefully designed national tests can help ensure a relatively wide distribution of test scores, many items in international assessments have been too difficult for students from less developed countries, resulting in restricted test score variance. This is reflected in the data presented in Figure 6.1 which is based on the results of a selection of countries that participated in the TIMSS 2003 study. The data show the percentage of Grade 8 students that reached levels or benchmarks of performance based on all students who took the test. Roughly three-quarters of those who took the mathematics test achieved the "low" international benchmark, one-half the" intermediate" benchmark, 23% the "high" benchmark, and 7% the "advanced" benchmark . In sharp contrast, 9% of Ghanaian students achieved the low benchmark and 2% the intermediate benchmark. Zero per cent achieved the "advanced" and "high" international benchmarks. Similarly on PISA 2003 , the limited utility of the assessment for internal policy making was underscored by the lack of test score variance in a number of participating countries; the majority of 15-year olds in Brazil, Indonesia and Tunisia scored below level 1. (It has been suggested that Level 2 be considered a minimum requirement for students entering the world of work and further education.)

Box 6.1

Percentage of Students Reaching TIMSS International Benchmarks1 in Mathematics, Grade 8: High and Low-Scoring Countries.

Countries	**Advanced**	**High**	**Intermediate**	**Low**
Singapore	44	77	93	99
Chinese Taipei	38	66	85	96
Rep of Korea	35	70	90	98

Internat. Average	**7**	**23**	**49**	**74**
Philippines	0	3	14	39
Bahrain	0	2	17	51
South Africa	0	2	6	10
Tunisia	0	1	15	55
Morocco	0	1	10	42
Botswana	0	1	7	32
Saudi Arabia	0	0	3	19
Ghana	0	0	2	9

1 Definitions used in TIMSS 2003:

Advanced: Students can organize information, make generalizations, solve non-routine problems, and draw and justify conclusions from data.

High: Students can apply their understanding and knowledge in a wide variety of relatively complex situations.

Intermediate: Students can apply basic mathematical knowledge in straightforward solutions.

Low: Students have some basic mathematical knowledge.

Source: Mullis et al (2004), p. 64

Publisher source : http://timss.bc.edu/PDF/t03_download/T03_M_Chap2.pdf

Eighth, a problem arises when the primary focus in reporting the results of an international assessment is on the ranking of countries in terms of the average scores of their students, usually the main interest of media. Rankings in themselves tell us nothing about the many factors that may underlie differences between countries in performance. Furthermore, rankings can be misleading when the statistical significance of mean differences in achievement is ignored while rankings can vary depending on the countries that participate, an important consideration when rankings over time are compared. Thus, for example, if the number of traditionally high achieving countries decreases and the number of traditionally low achieving countries increases, a country's ranking may increase without necessarily implying an improvement in achievement.

Ninth, poor performance in an international assessment (as well as in a national assessment) can carry with it some political risks for key officials associated with the delivery of education, including Ministers and Secretaries of Education. The risk is likely to be greater when the international rank of a country is lower than that of a traditional rival country. Some countries have actually collected data and refused to allow them to be included in between-country published comparisons when they saw the results. Such an approach possibly hinders the

development of a culture of rigorous assessment and evaluation within a country. Obtaining comparative data for neighboring countries or countries within a region would seem more appropriate than obtaining data for countries across the world that differ greatly in their level of socioeconomic development. Thus, ten Latin American and Caribbean countries jointly carried out an assessment of basic competencies in language and mathematics in 1997. The SACMEQ assessments in southern and eastern Africa by a network of ministries in the 1990s also allowed for international comparisons at a regional level.

Tenth, the primary purpose of both national and international assessments is to provide objective information on key aspects of the quality of the education system to enhance policy making. The likelihood that policy makers will use the results is likely to be enhanced when they have contributed to the design and implementation of the assessment. National assessments can be more easily tailored than international assessments to address the policy needs of key policy makers. For example, national policy makers can ensure that the assessment is directed at the grade level of most interest to them (e.g., the final year of primary school). Up to now, PISA has assessed the achievement levels of 15-year olds; in many developing countries, the majority of young people have left school by this age.

Eleventh, the demands of meeting deadlines may prove very difficult in countries that lack administrative personnel and have to cope with a poor communications infrastructure (see Box 6.1). Time allowed for carrying out various tasks (e.g., printing, distribution of booklets) associated with an international assessment , which may be deemed reasonable in developed countries, may be insufficient given the range of basic problems, including poor communication systems, that exist in many developing countries. Getting support staff can be difficult especially where Ministries of Education are slow in releasing funds to pay for services.

Box 6.1 South Africa's Experience with International Assessments

South Africa's experience with TIMSS and TIMSS-R underlines the problems facing implementers of international assessments. Deadlines imposed by organizers can be difficult, if not impossible, to meet in situations where there may be no mail or telephone services or funds for travel to schools. Other problems include lack of accurate population data on schools; poor management skills; insufficient attention to detail, especially in editing, coding, and data capture; lack of funding to support project workers; and difficulty in securing quality printing on time. Instructions to test administrators, for example to walk up and down the aisle, are obviously inappropriate when classrooms do not have an aisle.

Finally, in considering participation in an international assessment, countries with limited resources have to decide if the use of financial and human resources

can be justified when a great many other needs are demanding attention (e.g., the provision of teachers, running water, or electricity in schools). There are substantial costs associated with participation in an international study. A country participating in TIMSS for grade 8 was expected to pay US$ 40,000 plus all costs associated with printing, distribution, test administration, data entry, and scoring. It should also be recognized, of course, that national assessments also have considerable associated costs.

REFERENCES

Benveniste, L. (2002). The political structuration of assessment: Negotiating state power and legitimacy. Comparative Education Review, 46, 89–111.

Bernard, J-M (1999). Les enseignants du primaire dan cinq pays du Programme d'Analyse des Systèmes Educatifs de la CONFEMEN: Le rôle du maître dans le processus d'acquisition des élèves. Rapport réalisé pour le groupe de travail sur la profession enseignante, Section francophone de l'ADEA.

Carroll, D. (1996). The grade 3 and 5 assessment in Egypt. In P. Murphy et al (Eds.), National assessments: Testing the system (pp. 157–165). Washington DC: World Bank.

CONFEMEN (1999). Les faceurs de l'efficacité dans l'enseignement primaire: Les resultats due programme PASEC sur neuf pays d'Afrique et de l'Océan Indien. Dakar: Author.

Elley, (2005)

Elley, W.B. (Ed.). (1994). The IEA Study of Reading Literacy: Achievement and Instruction in Thirty-two School Systems. Oxford: Pergamon.

Ferrer, G. (2006). Educational assessment systems in Latin America: Current practice and future challenges. Washington DC: Partnership for Educational Revitalization in the Americas (PREAL).

Hanushek, E.A., & Kimko, D.D. (2000). Schooling, labor-force quality, and the growth of nations. American Economic Review, 90, 1184-1208.

Howie, S. (2005). System-level evaluation: Language and other factors affecting mathematics achievement. Prospects, 35, 175–186.

Ishino, T. (1995). Japan. In Performance standards in education. In search of quality, (pp. 149-161). Paris: OECD.

Johnson, E.G. (1992). The design of the National Assessment of Educational Progress. Journal of Educational Measurement, 29, 95-110.

Kellaghan, T. (2003). Local, national and international levels of system evaluation. Introduction. In T. Kellaghan & D.L. Stufflebeam (Eds.), International handbook of educational evaluation (pp. 873–882). Dordrecht: Kluwer Academic.

Kellaghan, T., & Greaney, V. (2001). Using assessment to improve

the quality of education. Paris: UNESCO: International Institute for Educational Planning.
Kulpoo, D., & Coustère, P. (1999). Developing national capacities for assessment and monitoring through effective partnerships. In Partnerships for capacity building and quality improvements in education: Papers from the ADEA 1997 biennial meting, Dakar. Paris: ADEA (Association for the Development of Education in Africa).
Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Chrostowski, S.J. (2004). TIMSS 2003 International Mathematics Report: Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Kennedy, A.M. (2003). PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools. Chestnut Hill MA: PIRLS International Study Center, Boston College.
Murimba, S. (2005a). The impact of the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ). Prospects, 35, 91–108
Murimba, S. (2005b). The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ): Mission approach and projects. Prospects, 35,75-89.
OECD (Organisation for Economic Co-operation and Development). (2003). The PISA 2003 assessment framework: Reading, mathematics, science and problem solving knowledge and skills. Paris: Author.
OECD. (2004). Learning for tomorrow's world. First results for PISA 2003. Paris: Author.
Postlethwaite, T. N. (2004). What do international assessment studies tell us about the quality of schools systems? Background paper for EFA Global Monitoring Report 2005.
Prakash, V., Gautam, S.K.S., & Bansal, I.K. (2000). Student achievement under MAS: Appraisal in Phase-II States. New Delhi: National Council of Educational Research and Training.
Raminez, F.O., Luo, X., Schofer, E., & Meyer, J.W. (2006). Student achievement and national economic growth. American Journal of Education, 113, 1-29.
Ravella, P. (2005b). Personal communication, March 2.
Štraus, M. (2005). International comparisons of student achievement as indicators for educational policy in Slovenia. Prospects, 35, 187–198.
Task Force on Education Reform in Central America (2000). Tomorrow is too late. http://thedialogue.org/publications/preal/tomorrow.pdf
UNESCO. (2000). The Dakar Framework for Action. Education for All:

Meeting our collective commitments. Paris: Author.
UNESCO. (2001). Technical report of the first international comparative study. Santiago, Chile: OREALC
UNESCO. (2002). EFA global monitoring report, 2002: Is the world on track? Paris: Author.
UNESCO. (2003). Monitoring Learning Achievement (MLA) Project. Update. Paris: Author.
UNESCO. (2005) Education for All global monitoring report 2005 – The Quality Imperative. Paris: Author.
U.S. National Center for Education Statistics. (2005). National Assessment of Educational Progress: The nation's report card, Reading 2005. Washington DC: Author.
U.S. National Center for Education Statistics. (2006) NAEP overview. http://nces.ed.gov/nationsreportcard/about/

НАЦИОНАЛНИ И ИНТЕРНАЦИОНАЛНИ ОЦЕНКИ НА УЧЕНИЧЕСКИ ПОСТИЖЕНИЯ

Резюме. Книгата представя основните характеристики на националните и международните оценявания на постиженията на учениците. И двата вида оценяване са се превърнали в популярно средство за измерване качеството на образованието от 1990 г. и 2000 г. насам. Този възход в популярността е повлиял и върху двата вида оценяване. Целите и основните черти на националното оценяване са описани в глава 2. Глави 3 и 4 описват причините, поради които трябва да се провежда национално външно оценяване и важните решения, които трябва да бъдат взети при планирането и дизайна на едно такова оценяване. Проблемите (както и грешките), които трябва да бъдат взимани под внимание при дизайна, провеждането, анализа, докладването на данните и изобщо необходимостта от национално оценяване, са описани в глава 5. Глава 6 се спира върху международните оценявания на постиженията на учениците, които в много голяма степен имат сходни стъпки на реализиране с националните оценявания (като изготвяне на извадка, администриране, начин на събиране на данните и методите за анализ).

Винсънт Грийн
Томас Келаган