DFID Department for International Development

Guidance Note

A DFID practice paper

National and international assessments of student achievement

Evidence on how successful schools are in transforming resources into student learning is essential to guide policy and management decisions regarding educational provision. Assessment of learning, especially in the foundational areas of language and mathematics, is needed at varying points in the educational careers of students.

Assessment entails measurement of learning, analysis to diagnose problems, and use of the findings to guide remedial action. An effective national assessment policy demands real political commitment to action based on the results, such as reallocation of resources, curriculum reform and/or re-orientation of teaching.

This Guidance Note describes the procedure known as a national assessment—or an international assessment if more than one education system is involved—that provides evidence on student learning. It identifies the issues to weigh up, explains their significance for educational improvement, and the factors which should be considered when discussing national assessment with government decision makers and civil society stakeholders.

National assessments of learning:

Summary

What is a national assessment?

Why conduct one?

What policy issues can it address?

What are the design considerations?

How should results be shared?

What is an international assessment?

What administrative and technical issues apply?

What is the cost?

How should the findings be used?

Guidance on options

Glossary

Further reading

A **national assessment** is a survey of schools and students (and sometimes teachers) that is designed to provide evidence, at the level of the education system, about students' achievements at a particular stage of education, in identified curriculum areas (e.g., reading or literacy, mathematics or numeracy, science).

An **international assessment** provides similar information for more than one education system but may not be sensitive to the characteristics of individual systems.

1. Summary

National assessments can play a critical role in demonstrating the efficacy or otherwise of all other investments in education. For a small proportion of education expenditure, they potentially play a unique role in determining value for money in the education sector. It follows that a national assessment should be undertaken on the basis of authentic political engagement with and commitment to the assessment programme, and a determination to allocate resources and reform education in the light of the findings. In a national assessment, the primary interest in collecting data is in what they tell us, when aggregated, about the **performance of the education system**, not the performance of individual students. Formative, continuous and summative classroom assessments of individual students are important aspects of schooling to help improve learning, but are not the subject of this Note.

There is no universally correct answer as to which options should be chosen by any one country at a particular historical moment with respect to each of the many aspects of assessment policy. Whilst the cost of assessment is relatively low (around 0.1% secondary education budget), it can be considerable in absolute terms (commonly several hundred thousand or over a million US dollars) when the opportunity costs of expenditure on other inputs are considered. That said, **the returns to national assessment increase**, as long as successive data sets are comparable and are formed into longitudinal series for analysis of trends.

A national assessment uses standardised instruments with an identified population or a representative probability sample of students in that population. Neither classroom assessments nor public examinations meet the criteria for national assessments. DFID is committed to working with governments and other development partners, to conduct **simple, sensible, sound and low cost assessments to improve learning outcomes** throughout the basic education cycle. Citizen-based assessments of learning such as ASER and Uwezo can complement national assessments and contribute to accountability. Under the right conditions, national and international assessments can form part of a quality improvement initiative in a developing country. They can provide transparent information about system performance that goes beyond input measures and reveals trends over time. Background information may allow results to be disaggregated for sufficiently large sub-populations, revealing relationships between achievement and student or school characteristics. Such information can guide investment of resources in low-performing geographical, administrative or curriculum areas, or for specific socioeconomic groups.

Government, in consultation with other stakeholders, **must specify at the outset the purpose** of the assessment and the research issues it is intended to address.

These can relate to quality, equity and provision. They will have implications for the design of the national assessment instrument and process, the technical focus of the assessment, and the population covered. Examples include whether to focus on curriculum content or life skills; identification of sub-populations to compare and contrast; and analysis of the effects of policy initiatives such as increased enrolment rates. Gauging the impact of key reforms or rapid change may demand annual or biannual assessments for a period; more typically three to four year intervals will suffice to assess systemic change.

Policy makers' and education managers' information needs should **determine decisions** on key design elements of a national assessment.

The considerations include:

- What curriculum area (eg, mathematics) or construct (eg, numeracy) to assess?
- At which grade(s) or stage(s) of education to assess?
- Whether to assess the whole school population at the chosen grade(s) or a probability sample of schools and students chosen to represent that population?
- How often to assess: annually or every three to four years?
- Whether policy/decision makers expect achievement to be described using raw scores, percentages, by curriculum domains/units, the proportion of students achieving curriculum attainment targets, or attaining specific proficiency levels?
- How, and to whom, should the findings of an assessment be communicated?

International assessments of student achievement—which may complement a national assessment or be the sole source of information about learning in a national education system—have been strongly promoted in recent years. An individual country's decision whether or not to participate in such an assessment requires a careful evaluation of its suitability: is the regional or international assessment framework appropriate for the country in question? Before embarking on an assessment, national managers of the process should satisfy themselves that the specific levels of technical skill (relating to decisions taken on test development, sampling and analysis) that will be required for its execution are available. If investment in capacity development or the employment of foreign technical assistance will be required, are sufficient resources available for that? DFID is committed to supporting partners to make effective use of national and international assessments, to provide a robust evidence base for improvement in learning outcomes over time. Temporary or strategic investment in national assessment capacity may be indicated.

Cost is another important consideration in deciding whether or not to carry out a national assessment, and whether to participate in an international assessment. Costs can vary considerably for an assessment depending on the range and depth of assessment instruments, whether the assessment involves all (or most) schools and students or is sample-based, whether reliable sub-national data or only national-level data are required, and the number and scope of follow-up activities planned.

Effective **use of assessment findings** includes applying the information gained to improve the quality of student learning. It follows that a commitment to successive assessment exercises over time is essential to fully reap the benefits of expenditure on assessment exercises. This demands institutionalisation of the assessment process, integration of assessment information into Education Management Information Systems, and alignment of national assessment to other elements of the education system such as community-based assessment initiatives. Political commitment to lead reform, evidence-based resource allocation and skilful change management—as well as the technical capacity in assessment—are integral to national assessment programmes.

There are significant caveats to consider if national assessment results are to be used to hold **teachers and schools to account for student performance**. A census-based assessment can provide diagnostic data on each school, help plan interventions, and inform communities/parents about individual school performance. But it can also create perverse incentives, significantly increase costs and drive distortions to the teaching/learning process.

A country with limited resources embarking on the development of a learning assessment system is faced with certain options and constraints. It is necessary to anticipate the limitations which inevitably result from choosing between alternatives, even in well-resourced international studies. **Guidance on preferred paths for developing countries** in typical circumstances is available at the end of this Note.

2. What is a national assessment?

The procedure used to assess student learning at the system level is variously referred to as a learning assessment, system assessment, assessment of learning outcomes, or national (or international) assessment—which is the term used in this Note. It is applied to a survey of schools and students that is designed to provide evidence about students' achievements in identified curriculum areas (eg, reading/literacy, mathematics/numeracy, science) for a clearly defined part of the education system (eg, fourth grade students or eleven-year olds).

In a national assessment,

1. achievement is assessed using standardised instruments, administration and scoring procedures;
2. assessment instruments are administered to an agreed-upon population of students or, more commonly, to a probability sample of students who are selected to be representative of the population;
3. individual student achievements are aggregated to the system level. Reliable data may also be obtained for subpopulations if samples are sufficiently large (eg, students categorised by the state/province in which they attend school; students attending private schools and students attending public schools);
4. background information, provided by participating students, teachers, and sometimes parents, is usually collected in questionnaires to provide insights into relationships between achievement and a variety of factors (eg, school and classroom resources and practices, student characteristics, family characteristics).

In **Uganda**, the National Assessment of Progress in Education (NAPE) has been conducted since 1996, as described by Kellaghan and Greaney (2008).

Test instruments were developed by the Uganda National Examinations Board (UNEB) and included standardised tests (English literacy and numeracy) and questionnaires for pupils, teachers and school principals.

A sample was drawn from each of the country's 14 administrative zones.

The proportion of students reaching each of four levels ('inadequate', 'basic', 'adequate' and 'advanced') were reported.

Information was disseminated through posters and user-friendly reports for teachers, principals, teacher educators and policy makers. Successive rounds of NAPE revealed a dip in pupils' mean scores and the proportion rated proficient in English and maths, following the rapid expansion of access to schooling due to introduction of universal primary education in 1997.

Recovery in overall performance was observed from 2003, despite wide disparities between districts driven in part by displacement of the population as the war in the north escalated. Gender differences by subject, and contrasting performance between urban and rural locations, were also noted.

NAPE has underpinned curriculum and timetabling changes to focus more on literacy and numeracy, as well as better classroom assessment practices. It helps head teachers, PTAs, school management committees, districts and national government make informed decisions about resource allocation. More teachers have been recruited and trained, their salaries improved, and damaged school infrastructure repaired in the north. (See also p15).

A national assessment differs from the kind of assessment that is found in everyday classrooms or lessons.

- Classroom assessment is an ongoing integral component of the teaching-learning process.
- The focus of classroom assessment is on the level of knowledge, skill or understanding of individual students in the classroom and the diagnosis of problems they may be encountering, with a view to deciding on the next instructional steps that need to be taken.
- The procedures used in classroom assessment are for the most part subjective, informal, immediate, and intuitive.

Classroom assessment does not meet any of the four criteria for a national assessment listed above: (1) standardised instruments and procedures; (2) administration to an entire student population or representative sample thereof; (3) system or sub-system level aggregation of results; and (4) systematic relation of performance data to background characteristics of students.

A national assessment also differs from public examinations. The latter may seem an attractive option for monitoring student achievement levels, since they are taken by large numbers of students and are already in place. It would require little added expenditure to carry out further analysis of results. However, they do not meet the second criterion (universal/representative application) or fourth criterion (background data collection) for a national assessment listed above, while interpretation of aggregated data (the third criterion) is problematic.

- The information public examinations provide is limited to students who survive in the system and who take the examinations at the end of a phase of education. However, policy makers should, for political, economic and social reasons, be interested not only in this successful elite, but also in less successful students. They should also be interested in achievement levels at earlier stages of education when interventions can be effective in improving outcomes.
- Since they are often used for selection, a major focus in designing public examinations is to discriminate between candidates at critical points on the achievement scale and not on the achievements of the considerable number that fall below them.
- When candidates are given a choice in the subjects in which they take examinations (and may also have a choice in the questions they respond to), interpretation of what aggregation of individual student results means is highly problematic.
- The content of examinations, as well as the characteristics of students, change from year to year, limiting the inferences that can be made from comparisons over time. There is evidence from a number of countries that the proportion of examination candidates being awarded high grades has increased from year to year, even though standardised measures of student achievement show no improvement. For example, research at the Curriculum, Evaluation and Measurement Centre of the University of Durham estimated that in England, when matched for ability, GCE A-level candidates in 2007 achieved B grades whereas their peers of 1997 received only C grades.

The last two decades have witnessed **rapid growth in the number of countries** carrying out system-level assessments of student achievement. Since the Dakar conference in 2000, almost 40% of countries in Sub Saharan Africa have conducted at least one national assessment, compared to about 25% prior to 2000. The region, however, together with central Asia, still exhibits the lowest level of system-level assessment.

3. Why conduct a national assessment?

Collecting and publishing statistics on quantifiable inputs to the education system—eg, physical facilities, student enrolments and teacher-pupil ratios—does not tell us if students have benefitted from the inputs as reflected in their learning. Rigorous and periodic assessments of student learning are necessary to provide **evidence** of the extent to which the considerable amount of money that is spent on education does, in fact, result in student learning.

If policy makers and education managers do not know how successful (or unsuccessful) schools are in **transforming resources into student learning,** they risk maintaining sub-optimal education environments.

When compared with total expenditure on education, a national assessment is likely to be a relatively inexpensive supplement to reform efforts to improve learning. However, the absolute **costs may be significant**, especially in countries where the non-salary budget is small. The cost of a national assessment will be sufficient to buy many textbooks or employ more teachers, making it unattractive to policy makers.

The value of national assessment increases as the time series of comparable data builds up. Therefore, the decision to allocate funds should be considered as an on-going commitment over the medium and long term, not a one-off expenditure.

Notwithstanding the above, there are **high opportunity costs in not undertaking assessments**. Apart from the fact that assessments tell us how successful the education system is in promoting student learning, without repeated measurement over time we do not have evidence of trends that can be used to guide policy and further investment.

Policy makers and education managers need the empirical data that assessments provide in making decisions regarding the allocation of resources. For example, a national assessment can identify areas of the curriculum in which a considerable proportion of students are underachieving. Furthermore, **underachievement** may be found to be associated with specific factors such as location (eg, schools in rural areas) or type of school. Ensuing action may involve the provision of inservice courses for teachers or of additional resources to schools in specific categories.

Apart from the direct benefit of having a reliable 'snapshot' of learning outcomes, involvement in national assessments **develops capacity** in a number of critical areas, including the analysis of data and the dissemination of findings. This capacity may subsequently be used to improve the national education quality assurance system.

National assessments provide data which can be used for **secondary analysis** by universities, research institutions, and regional education authorities.

If the results of assessments are effectively communicated and disseminated, they will raise the profile of education and **stimulate national debate**, garnering support for education policy reform and additional investment. Perhaps the best example of this is in the USA where the campaign to raise awareness of the findings of the National Assessment of Educational Progress has resulted in the survey being widely regarded by the public as The Nation's Report Card.

For all these reasons, the collection of information on student achievement at the system level is considered to be an essential component of an education management information system (EMIS) in addressing the need to base policy and practice on concrete evidence about the needs of students and schools.

In **Germany**, PISA data, used to carry out national-level analyses, revealed major differences in mean achievement between its 16 states. Results, which were interpreted as providing evidence of social inequity, gave rise to widespread media, public, and political reaction. Ensuing reforms included the introduction of common standards across states, the preparation of an annual report on education in the country, and an increase in the number of hours of schooling.

4. What policy issues can a national assessment address?

The findings of a national assessment are more likely to be used if a ministry of education makes clear at the outset the **purpose** of the assessment. The key research issues that it expects the study to address should be specified at the outset, preferably in consultation with other stakeholders. Research issues can be categorised as relating to quality, equity and provision.

Issues relating to quality

A national assessment can provide information about the quality of student learning with reference to national statements of educational standards, the implementation of the curriculum, public perceptions about what students should be able to do, and whether or not students are properly prepared for future life. The interest expressed by policy/decision makers will have implications for the design and content of the assessment instrument (eg, does it focus on curriculum content or does it attempt to identify life skills?)

Assessment data can be used to monitor change in achievement over time. Reliable data are necessary if educational authorities are to answer the question, "Is the quality of our education system, in terms of learning outcomes, improving?"

Table 1: Questions that national assessment can address

Examples of questions on quality that a national assessment could address…	…and the implications for design that follow
What proportion of children of primary school leaving age are functionally numerate?	Measure children's performance using tasks in tests based on mathematical constructs for the appropriate developmental stage
What proportion of children of primary school leaving age have attained mastery level competence in mathematics?	Measure children's performance using tasks in tests based on primary mathematics curriculum learning objectives
What proportion of children at the beginning of grade 3 can read? How has that changed by the time they reach grade 6?	Measure children's performance using reading tests based on phonemic awareness, phonics, fluency, vocabulary and comprehension appropriate to their grade, with a probability sample of grade 3 cohort followed up at grade 6.
Which schools are most/least effective at converting resources into learning outcomes?	Assess all children of given stage/age at all schools nationally
Which types of schools are most/least effective at converting resources into learning outcomes?	Assess a sufficiently large probability sample of children of given stage/age from each category of school of interest
How are children at a given stage performing when compared to their peers in other countries in the region?	Join a relevant and comparable regional international assessment protocol
How are children at a given stage performing when compared to their peers in previous cohorts?	Build up a time series of national assessment data using tasks and tests which are of comparable standard over time
What changes should a primary school teacher make to reflect children's different learning styles and abilities in class?	Do not use national assessment: develop a suitable formative classroom assessment protocol
Which children should be selected for scarce secondary school scholarships?	Do not use national assessment: develop a suitable summative assessment process and selection procedure towards the end of the primary stage

Issues relating to equity

A national assessment can help determine if the education system is underserving particular groups of students as evidenced in differences in achievement related to:

- gender;
- location (administrative division, urban-rural);
- ethnic or language group membership;
- socioeconomic group (eg, low-income families, scheduled tribes and castes in India);
- school governance (government, government-aided, private).

Assessment data can indicate if disadvantaged and underperforming groups improve over time. This is particularly important when interventions have been made to improve the learning outcomes of a specific group.

Issues relating to provision

A national assessment can provide empirical evidence about a variety of aspects of provision if relevant data are collected in conjunction with student achievement data. These might include

- factors associated with achievement (e.g., school resources; the amount of time spent teaching a curriculum area; teachers' level of training/qualifications; students' home circumstances);
- the effects of restructuring or decentralisation of the education system;
- the effects of curriculum reform;
- the effects of increasing student enrolment rates, especially when this is the result of government policy;
- monitoring grade repetition and its impact on teaching and learning.

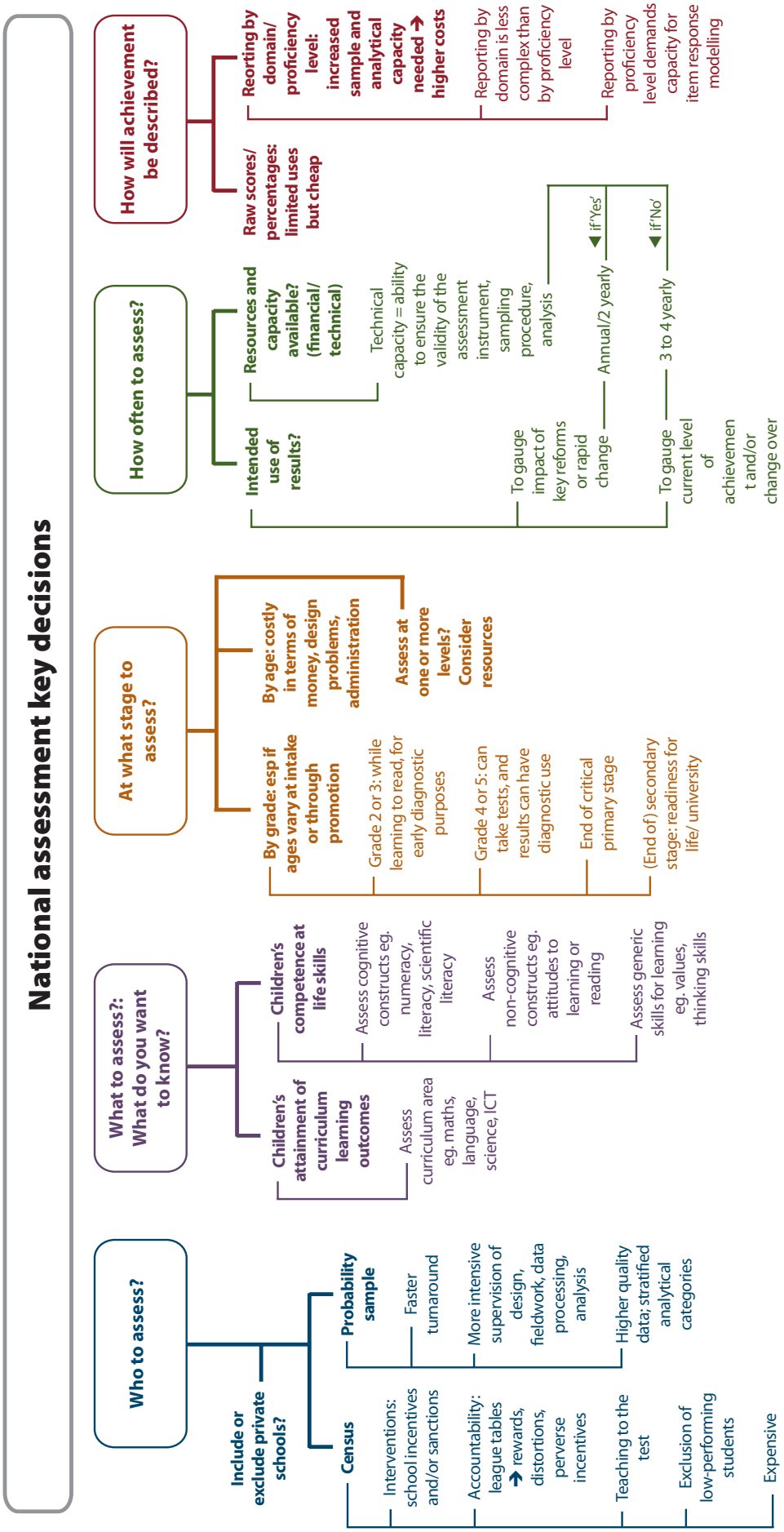
Assessments which relate teachers' competence to students' achievements are especially noteworthy given contemporary concerns with payroll expenditure and value for money. In **Vietnam** both teachers and pupils took Grade 5 tests in mathematics and reading, with results calibrated to the same continuum. Overall, the tests proved appropriate for the pupils and easy for the teachers, but with some overlap (some pupils performed better than some teachers). Most tellingly in terms of policy on education provision, the exercise revealed significant differences between provinces, with a linear relationship between teacher competence and student performance in each province. In **Kwara State, Nigeria**, all 19,000 primary and junior secondary teachers were assessed on basic (primary grade 4) literacy and numeracy skills, and their ability to apply these to everyday teaching tasks such as lesson planning. The report's 'startling' finding of 0.4% of teachers demonstrating the minimum competency level (a score of 80% on primary 4th grade subject knowledge), taken in conjunction with a separate systematic assessment exercise to monitor pupils' learning achievement, provide a unique insight into the massive constraints on improving children's learning outcomes. To the state government's credit, they have been used to drive the reform agenda, with a tailor-made support programme for improvement of classroom teaching launched on the back of the assessment findings, and with the processes being replicated in certain other Nigerian states.

It should be noted that while such assessments provide objectively verifiable evidence about learning outcomes, the findings have to be interpreted 'intelligently'. For example, when government policy significantly increases enrolment from disadvantaged groups, the outcome may be a deterioration in average standards—a potentially embarrassing finding for politicians. It is therefore incumbent on personnel providing technical assistance to national assessments to be alive to the political sensitivities of the work. Results may take months or even years to reach the public domain, if the domestic political agenda so dictates.

All parties should be clear about the purpose of a national assessment and the uses to which the data will be put, before embarking on one. Assessment merely to indicate a government's commitment to achieving Education for All goals, or to move into line with international expectations, is unlikely to result in findings being given serious consideration in policy making, revised resource allocations, or being applied by managers to improve the education system. Assessment should be thought of as an inherently political exercise as well as a technical one.

5. What are the design decisions for a national assessment?

The design of a national assessment requires a number of decisions which will be made in light of the information needs of policy makers and education managers. This section can be read in conjunction with the mind map that follows.



Total population (census) or sample assessment? *Issues and options*

A national assessment in which all (or nearly all) schools and students at a specific grade or age level participate is termed census- or population-based. By contrast, a national assessment may collect information in a probability sample of schools and students which are chosen to be representative of the population at a specified grade or age level (sample-based).

Using sophisticated sampling techniques, the absolute size of the requisite sample may be relatively small, even for countries with extremely large populations. For example, the Russian Federation has a population of 143 million. However, in TIMSS 2007, a representative sample was achieved by testing 4,659 Grade 4 students in just 210 schools. A **sample-based assessment has three major advantages** over a census-based assessment:

- Since fewer schools are included in the test administration and fewer students are tested, costs for administration, scoring student responses, data entry and data processing are less. In a large country, the difference is huge.
- Turn-around time is faster as less time is required for data preparation and analysis.
- More intense supervision of fieldwork and of data preparation is possible, thereby ensuring higher quality of data.

In most countries, national assessments are sample-based. In some, both census and sample-based assessments are carried out for specific purposes. In **France**, all students are assessed at the beginning of the first year of lower secondary school, and diagnostic information on each student's progress, strengths, and weaknesses is sent to schools. A sample-based assessment is administered at the end of lower secondary schooling.
Both types of assessment have also been carried out in **Australia, Costa Rica, Cuba, Mexico**, and **Uruguay**.

A major difference between a census and a sample-based assessment is that census-based assessments provide information about **all schools**. This may be used to identify poorly performing schools, which may, in turn, be followed by some form of intervention. Census-based assessments also frequently form part of an accountability system in which schools (and teachers) are **held accountable** for their students' performance. (Notable examples include **England** and individual states in the **United States**.) In some cases, the performance of schools is published in league tables, and sanctions (incentives or penalties) are attached to performance for teachers and even students.

When used to hold schools accountable in such ways, a census-based assessment becomes a "**high stakes**" operation which has been found to have a number of adverse effects on the quality of students' education. Negative consequences have been found to include the following:

- Teachers tend to react by aligning their teaching to the knowledge and skills assessed in the test, neglecting curriculum areas not assessed. This highlights the importance of aligning the assessment protocol to core curriculum objectives. Advocates claim a benign influence of assessment on adherence to the curriculum, and argue that it can help leverage compliance with national standards.
- Teaching tends to emphasise rote memorisation, routine drilling and accumulation of factual knowledge rather than the acquisition of higher-order general reasoning and problem-solving skills. Assessments of creativity can be devised, but are atypical in national assessment programmes, and are particularly difficult to support where teachers' knowledge and pedagogical skills are impoverished.
- Teachers focus their efforts on pupils who are just below critical thresholds to help them make the grade, which, in turn, will make their school look good. (This phenomenon is well documented in the US where such students are termed 'bubble kids'.) The introduction of multiple cut-offs could go some way towards mitigating this effect.
- Low achieving and disadvantaged students tend to drop out of school, and there is a disincentive for schools to work to retain them if the school's performance is depressed by doing so.

In **Chile**, which has a high-stakes census-based assessment, both positive and negative consequences have at times resulted. Schools are ranked in terms of their performance nationally, and with respect to other schools in the same socioeconomic stratum to estimate 'value added'. The best performing schools are identified and their teachers financially rewarded. Schools in the bottom 10% of ranks receive assistance (materials, advice manuals) to improve teaching. However, teaching, learning, and parents' choice of school are strongly—perhaps unduly—influenced by the ranking based on the assessment. The existence of perverse incentives has led some teachers to misrepresent information about their schools to obtain resources.

Total population (census) or sample assessment? *Decisions*

Policy/decision makers should consider the advantages and disadvantages of sample- and census-based assessments.

A **sample-based assessment** has several advantages (including cost) if all that is required is information about the achievements of students in the population (and in subpopulations) and factors associated with achievement. It will not, however, provide information on individual schools.

A **census-based assessment** will be required if policy/decision makers seek information about all schools in the education system

- to provide diagnostic data to each school;
- to plan interventions or decide on resource allocation in individual schools that are identified as experiencing problems;
- to inform parents and communities about the performance of individual schools;
- to institute an accountability system.

Policy/decision makers should be aware of the **likely negative consequences** if they are considering attaching sanctions to performance in a census-based assessment.

Whether an assessment is based on the total population or a sample, a decision to **include or exclude private schools** has to be made.

What curriculum area or construct to assess? *Issues and options*

All national assessments measure the cognitive outcomes of instruction or scholastic skills. Language/literacy and/or mathematics/numeracy feature in most assessments, a recognition of the important foundation they provide for all school learning. Children who do not learn to read in the first few grades are likely to repeat grades and to drop out at an early stage. In some assessments, knowledge of other curriculum areas (eg, science, social studies) is included. A few assessments have collected information on non-cognitive outcomes (eg, attitudes to learning, attitudes to reading, values, self-concept). Of particular interest, though not widely implemented, is the measurement of students' generic skills that contribute to the process of learning (eg, self-regulation, self-confidence, engagement, motivation, and the Personal, Learning, and Thinking Skills of the English national curriculum).

What curriculum area or construct to assess? *Decisions*

Policy/decision makers have to decide on the curriculum area or construct to be assessed. For example, the assessment may focus on achievement related to the learning objectives (ie, content) of the national curriculum for mathematics, language, or science. Alternatively, the assessment could look at carefully designed constructs such as numeracy, reading literacy or even 'scientific literacy'. As literacy and numeracy are key areas of achievement, priority should be afforded to these, especially when the assessment targets young learners.

Policy/decision makers may have particular concerns about other areas of achievement (eg, science or ICT in secondary schools) that would lead them to propose their inclusion in an assessment. The assessment of these areas, however, is likely to be less frequent than assessments of literacy and numeracy.

At what stage to assess? *Issues and options*

Target grades for national assessments vary from country to country, ranging from grade 1 (in Uruguay) to grade 12 (in the United States). Few countries conduct large-scale assessments before grade 3 using paper-and-pencil tests, as these present problems for pupils who may not be skilled in reading, writing or in following instructions. However, interest in developing assessment procedures that will identify problems at an early stage, as well as ones that will be appropriate for assessing children who are not attending school (eg, ASER community-based assessments in India, and Uwezo inclusive assessments in East Africa), remains strong. An assessment instrument (prePIRLS) is being developed for administration in 2011, to test more basic skills than are currently assessed in the international study PIRLS.

The FTI places learning outcomes at the centre of its agenda. The two reading skill indicators in its Indicative Framework are:

- Proportion of students who, after two years of schooling, demonstrate sufficient reading fluency and comprehension to "read to learn".
- Proportion of students who are able to read with comprehension, according to their countries' curricular goals, by the end of primary school.

The **Early Grade Reading Assessment** (EGRA) adopts a more radical approach and uses an oral assessment protocol designed to measure the most basic foundation skills for literacy acquisition in grades 1, 2, or 3 (phonics, phonemic awareness, vocabulary, fluency, comprehension). EGRA has been used in about 70 countries to provide national or system-level diagnostic information about children's early learning, as well as to support classroom-based assessment and in programme evaluations. As the assessment is individually administered, requiring about 15 minutes per child, its use to provide national-level data would be expensive. Also relevant is the fact that EGRA cannot be compared across languages (even within the same country), and that some elements of the EGRA test battery cannot be applied for some languages.

In all countries where national surveys are carried out, assessments take place in the primary school grades. In many, they are also conducted at some point at secondary school level, usually in the lower-secondary grades. Information at **both levels can be valuable**. At primary level, assessments can, at an early point, identify deficiencies that might underlie difficulties at a later stage. In some cases, policy makers have opted for an assessment at the end of primary school following discontinuation of a public examination at that point.

An important decision regarding the population to be assessed is whether it will be defined by age or by grade level. **Most national assessments opt for grade**. If, because of differing ages of entry to school and/or policies of non-promotion, students of a similar age are not concentrated in the same grade, choosing a population on the basis of age may be considered. However, this would be disruptive in schools as it would require students from several grades to take tests at the same time. It would also be very difficult to identify appropriate test content for such students and would be more costly to administer.

It also assumes the existence of tolerably accurate data regarding children's date of birth.

At what stage to assess? *Decisions*

Choice of grade (or age) level for a national assessment will depend on the point at which information would seem most useful. This could be

- during the early grades of primary schooling (grade 2 or 3) when pupils are acquiring basic skills and have not made the transition from learning to read to reading to learn and when an assessment might identify deficiencies that would be likely to create problems at a later point;
- sometime during the course of primary education (grade 4 or 5) when it is still not too late to identify early learning difficulties, but pupils are more accustomed to taking tests;
- at the end of the critical primary school stage;
- sometime during the course of secondary education, most likely about the point of the end of compulsory education, to obtain evidence on how well students are prepared for further education or for life after school.

Policy/decision makers should be aware that **choice of age rather than grade for an assessment presents problems** for the design of tests and for administration in schools.

A decision to hold an assessment at only **one grade level or at more than one grade level** will usually depend on the resources available.

How often to assess? *Issues and options*

In deciding on the **frequency** with which data on student learning will be obtained, the main considerations are the intended use of results, and expense. If the major objectives of a sample-based assessment are to obtain empirical information on current levels of achievement for the education system as a whole and for subpopulations, and to monitor possible changes over time, an assessment every three or four years is adequate, as achievement levels change very slowly.

There may, however, be particular circumstances that indicate the need to obtain information over a shorter time period. For example, if reforms are being introduced, or if participation rates in the education system are expanding rapidly (eg, following the implementation of a policy of providing Universal Primary Education), policy makers may wish to monitor the effect on a more frequent basis.

The amount of money available for an assessment is a major consideration in determining the frequency with which it is carried out. Cost, even in a sample-based assessment, is considerable, especially if separate information is required for administrative units (provinces, states) in the system. International assessments are more expensive than regional ones.

How often to assess? *Decisions*

Specific justification in terms of information use is needed, if assessment of student achievement levels is to be conducted more frequently than every three or four years. Such reasons could include monitoring the impact of reforms or rapidly changing participation rates.

The availability of resources, including money and human capacity, to conduct high quality assessments more frequently (eg, every year) is a major consideration. Options and cost drivers have to be assessed in national context. All forms of national assessment of learning represent a tiny proportion of total expenditure—typically in the range of 0.01% of the education budget, and 1% of alternative learning improvement interventions such as raising teachers' salaries or reducing class sizes.

How will achievement be described? *Issues and options*

A **variety of methods** have been used to report the findings of a national assessment. Some describe student achievement simply in terms of raw or percentage scores. These, however, provide little indication of what students can and cannot do. More sophisticated methods of analysis and presentation are required to address this issue.

Scores presented in their raw form as **percentage correct or as a linear transformation** (eg, with a mean of 500 and a standard deviation of 100), are appropriate in analyses

- to compare the performances of subpopulations;
- to estimate the percentage of examinees scoring at or below a given score;
- to relate achievement to background factors;
- to monitor average performance over time.

Greater insight into student achievement is provided if information is available for separate domains. If a test provides an adequate representation of aspects of the curriculum or construct, items can be grouped into **curriculum units or domains**, allowing student achievement to be reported in terms of performance in each domain. For example, reading comprehension items might be classified by ability to retrieve information from a text, to make straightforward inferences from a text, and to interpret, integrate and evaluate text information. Mathematics performance might be reported for number, measurement, shape and space, and data representation.

Student performance may be reported as the **proportion of students achieving attainment targets in the curriculum**, if the curriculum is structured in terms of the level of performance expected of students at particular stages (eg, the national curriculum in England).

An alternative approach to describing achievement involves reporting performance in terms of **proficiency levels** which describe what students can do at varying points in the achievement range. Since proficiency levels describe student performance in terms of a **hierarchy of knowledge and skills**, students at a particular level would be expected to be able to do what was required at lower levels, but not at higher levels. The number of levels varies from study to study. Some identify two levels (mastery/non-mastery), although in this case the definition of 'mastery' is often arbitrary (eg, a predetermined percentage correct score on a test) rather than in terms of defined proficiency. In some assessments, three levels (basic/proficient/advanced) are specified, as seen in the USA's NAEP grade 4 maths assessment (Annex 1). Other more detailed descriptions, such as the Vietnamese national assessment of reading skills, involve five or six levels (Annex 2). The percentage of students scoring at each level is calculated and reported. More resources and examples of levels, key stages and indicators of work standards for each level are available on the website of the National Curriculum for England.

Assessment reports routinely provide data on the **distribution of achievement**. Analyses that calculate the extent that schools do or do not perform at comparable levels have attracted considerable attention in some studies. Large differences between schools may reflect the characteristics of students in a school and of their communities, selectivity of the education system, and/or differential school effectiveness. Education systems in which the national level of achievement is low tend to exhibit large differences between schools in their achievements.

How will achievement be described? *Decisions*

Policy/decision makers should be made aware of the variety of approaches that exist when deciding how the achievements of students should be represented. In particular, they should be aware of the limitations of reporting performance only in terms of mean score correct. If their preference is for reporting by domain or by proficiency level, they should be aware that this will require a wide sampling of student achievement (involving increased cost) and the services of expert curriculum/construct analysts. Reporting by domain is less complex than reporting by proficiency level which requires a person with advanced statistical skills (eg, the ability to use item response modelling).

Policy/decision makers should indicate in advance what information they expect to obtain about the distribution of achievement in the population. This will depend in turn on the uses to which the assessment findings will be put.

Similarities and differences between assessment regimes

An examination of the assessment regimes of three countries (Chile, Sri Lanka, Uganda) reveals similarities (all assess language/literacy and mathematics/numeracy in the primary school grades) but also many differences (in the range of grades assessed; in basing the assessment on all schools or on a sample; in the inclusion or exclusion of private schools; and in the frequency of assessment).

National Assessment in Chile

System assessment in Chile goes back to 1978 when the ministry of education asked the Pontificia Universidad Católica de Chile to design and implement an information system for education. Since then, the history of assessment efforts has been erratic and at times controversial. Administration of the assessment has been taken over by the ministry of education.

Construct/curriculum area assessed: Spanish reading and writing, mathematics, social science, natural science.

Other areas assessed: Students' self-concepts.

Targets: Grades 4, 8, and 10 in public and private schools.

Sample or census: Census: Spanish and Mathematics for all students in the targetted grades; social science and natural science for approximately 10% of students.

Frequency: Annual (since 1996).

Background data: Questionnaires for principals, teachers, and parents (one year only).

Reporting: Each school receives an individual report on performance. Schools are ranked in terms of their performance nationally and with respect to other schools in the same socioeconomic stratum. The best performing schools are identified and their teachers financially rewarded. Schools in the bottom 10% of ranks receive assistance (materials, advice manuals) to improve teaching.

Impact: Results are widely publicised in the media and used extensively in policy discussions. Data from the assessment were used to guide decentralisation of the management of the education system.

Issues: Schools and teachers are considered accountable for students' learning based on their performance in the assessment. Teaching, learning, and parents' choice of school are being driven by the ranking based on the assessment. Some teachers misrepresented information about their schools to obtain resources.

Participation in international assessments: Chile has participated in TIMSS, PISA and LLECE.

National Assessment in Sri Lanka

The first national assessment in Sri Lanka was carried out by the National Institute of Education in 1994 (at grade 5) in conjunction with the Monitoring Learning Achievement (MLA) project organised by UNESCO/ UNICEF. With the setting up of the National Education Research and Evaluation Centre (NEREC) at the University of Colombo with funds from the World Bank, national assessment activity was placed on a firmer footing. NEREC carried out assessments since 2003 at grades 4, 8, and 10. Since 2008, the World Bank has supported the development of capacity to carry out assessment at the Open University, which conducted an assessment at grade 10 in 2009.

Construct/curriculum area assessed: First language (Sinhala, Tamil), mathematics and English (grade 4); first language, mathematics, science and technology (grades 8 and 10); English language (grade 10).

Targets: Grades 4, 8, 10.

Sample or census: Samples in all nine provinces.

Frequency: Grade 4: 2003, 2007, 2009; Grades 8 and 10: 2005, 2008; Grade 10: 2009

Background data: Questionnaire data collected from students, school principals, teachers, parents, zonal education officers, and inservice advisers.

Reporting: Results are presented for percentage of students who have 'mastered' the subject area, with 'mastery' being based on an arbitrary score of 80%.

Provinces and districts were rank-ordered in each subject area. Mean achievement scores were compared for school types, location, gender, and level of teacher training.

Impact: A variety of effects has been attributed to the findings of national assessments, including impact on curriculum reform and on teacher training programmes.

Issues: Private schools are not included in the assessment. The large improvement in performance between 2003 and 2007 in the grade 4 assessment raised questions about the equivalence of samples and of the conditions under which tests were administered.

Participation in international studies. Items from TIMSS were included in the grade 4 national assessment in 2007.

National Assessment in Uganda

Several national assessments have been carried out in Uganda since the 1990s. At one stage, three were in progress, one of which was indigenous and operated by the Uganda National Examinations Board (UNEB), the other two, administered by the Ministry of Education and Sports, had international connections (the Monitoring Learning Achievement project and SACMEQ). UNEB, which is now established as the primary agency, has developed a specialised unit to administer assessments.

Construct/curriculum area assessed: Literacy (English), numeracy, social studies, natural science.

Targets: Grades 3 and 6. A sample of teachers is also assessed.

Census or sample: Sample.

Frequency: Literacy and numeracy have been assessed annually since 2003. Oral skills in English are assessed every three years.

Background data: Questionnaires completed by principals, teachers and students (every three years).

Reporting: Student performance is categorised as 'adequate' ('desired'), basic or inadequate. Cut-scores for each level were set by expert panels. Results are also reported for pupil age, school location (urban/rural), geographic region and zone. Results are presented in a user-friendly way on posters that are distributed to schools. The technical report includes recommendations to address identified shortcomings.

Impact: The Uganda national assessment has had a strong focus on using findings to improve classroom teaching. Findings are reviewed in local workshops and approaches, including the improvement of classroom-based assessment, to address weaknesses in student achievement identified in the assessment. Findings have also been used to inform curriculum reform and teacher training.

Issues: Sampling relies largely on EMIS records which seem to be out of date. The result is a relatively poor response rate, especially in private schools (about 50%). A sharp improvement in performance on the numeracy test in 2008, and maintained in 2009, led to questions about the validity of the data.

Participation in international studies: SACMEQ.

6. How should results be shared?

It is important to anticipate **how, and to whom, the findings of an assessment will be communicated** when designing the assessment. These factors will affect decisions on the analyses that will be carried out, and in making budgetary provision.

While politicians may sometimes resist revealing the findings of an assessment, the long-term **advantages of an open information system** are likely to outweigh short-term disadvantages. Where findings have been widely disseminated, they have raised public awareness, making education an issue on the public agenda. If the assessment has been conducted with external technical assistance, serious consideration should be given to how to assist the host government address any shortcomings revealed. Anticipation and active management of media coverage is also required. In this way, a potentially harmful event can be used to spearhead a constructive reform effort.

A **detailed report** of an assessment, describing procedures followed and results, should always be prepared. This usually contains technical information, though in some cases a separate technical report is prepared. The technical information should be sufficient to allow members of the research community to evaluate the study critically. It also acts as a record of the activities of the assessment which is needed to implement future cycles of an assessment.

A number of additional means of communicating the findings of an assessment in non-technical and **easy-to-read language** will be required, to meet the needs of the many stakeholders who have an interest in the findings. A variety of reports are listed in Table 2, if sufficient resources are available. Kellaghan, Greaney and Murray (2009) can be consulted for further reading on this topic.

Table 2. Reports Following a National Assessment

Main/technical report	The primary source of information about an assessment. It describes the content of the assessment, its objectives, the framework that guided design, procedures followed, a description of achievement, correlates of achievement, and change in achievement over time (if appropriate data are available from a number of assessments).
Briefings (written and/ or oral) for minister and senior policy personnel	Capture the main findings of an assessment in a concise form, and identify possible implications.
Summary reports	Contain descriptions of the main findings in nontechnical terms for community leaders, employers and business leaders, the general public.
Reports for educationists	Present findings that are of particular relevance to teachers, teacher trainers, curriculum developers.
Press releases and press conferences	Present summary findings and key messages for radio, television, printed media.
Thematic reports	Explore aspects of the findings of an assessment related to a specific theme that is not addressed in detail in the main report (eg, gender, students in disadvantaged circumstances).
Web-based dissemination, knowledge platforms and downloadable data sites	Supports further investigation by the national and international research community, media, education profession, and general public.

7. What is an international assessment of student achievement?

An international assessment of student achievement is **similar** in many ways to a national assessment. It conforms to the definition used in this Note: a survey of schools and students that is designed to provide evidence at the level of the education system, about students' achievements at a particular stage of education, in identified curriculum areas. Both allow for tracking change over time, if conducted in successive cycles on a comparable basis. National and international assessment exercises have a broadly similar approach and share similar procedures in terms of instrument construction, sampling, scoring and analysis. Frequently, the results of an international assessment are used by individual countries to carry out their own within-country analyses.

The most obvious **difference** between a national and international assessment is that the latter provides information about an education system in relation to one or more other systems. However, the information it provides may be of limited value if the assessment framework is inappropriate for a country. In this context, it is worth noting that a developing country joining an international study may not be able to influence the assessment framework.

Many of the countries that participate in an international study also run their own national studies. It is not always clear whether this is the result of serious consideration of the benefits that each provides, and how they might complement each other, or is a consequence of poor planning and lack of **co-ordination** between decision makers in ministries of education.

Table 3. Benefits and risks of joining international assessments

Potential benefits of joining international assessment exercises	Potential risks of joining international assessment exercises
Adherence to high technical standards of assessment design, instrumentation, sampling, administration, analysis and reporting	Criticism of the cost of participation, particularly in view of the need to commit to successive rounds if the initial investment is to be worthwhile
Development of indigenous capacity to meet international standards of assessment practice	Disaffection with the international exercise if its assessment framework is of limited relevance and responsiveness to the country joining
High degree of transparency in dissemination of the results; political gains if performance is found to be relatively good compared with peers	Unfavourable comparison of results with neighbours and peers—with attendant political consequences
Positive effects of: driving up performance from diagnostic application of results; exposing education system to external scrutiny; and tracking impact of certain interventions/reforms over time	International assessment exercises should not be expected to deliver the accountability outcomes that national census assessment exercises provide
Opportunity to 'version' survey instruments of international standard, ie, to adapt them to the national language and context	Failure to fully adapt the survey instruments to the national context. 'Versioning' goes beyond translation, to ensuring that literacy texts (in particular) are suited to the children's educational experience and sociolinguistic background.

International assessments of student achievement fall into one of two categories: **global assessments** or **regional assessments.**

Global assessments: *Options*

International assessments that are carried out in countries throughout the world may be considered global. Currently, there are four such assessments.

- prePIRLS, organised by the International Association for the Evaluation of Educational Achievement (IEA), is being designed, following the PIRLS framework, for countries in which most students at grade 4 are earlier in the process of learning to read than grade 4 children from those countries that participated in PIRLS. prePIRLS is designed to test foundational skills for reading, such as recognising words and phrases, and making straightforward inferences.
- *Progress in International Reading Literacy Study (PIRLS)* is organised and co-ordinated by the IEA. PIRLS measures trends in children's reading literacy achievement, policy and practices related to literacy. Newly developed texts and questions for 2011 will be presented together with items from 2001 and 2006. The study includes assessment of relatively complex reading comprehension tasks such as may be found in school subjects.
- *Trends in International Mathematics and Science Study (TIMSS)* is also organised and co-ordinated by IEA. TIMSS 2011 comprises the fifth assessment in a series begun in 1995. It will collect data in mathematics and science at fourth and eighth grades.
- *Programme for International Student Assessment (PISA)* is organised and co-ordinated by the Organisation for Economic Co-operation and Development (OECD). PISA 2009 surveyed 15-year-olds in 65 industrialised countries/economies. Every three years, PISA assesses how far students near the end of compulsory education have acquired some of the knowledge and skills essential for full participation in society.

Table 4. Global Studies of Student Achievement

	Construct/ Curriculum Area Assessed	Age/ Grade	Frequency	Participating Countries
PrePIRLS	Reading	Grade 4	Every 5 years	Commencing in 2011
PIRLS	Reading	Grade 4	Every 5 years	40 mainly industrial countries (2006) 57 scheduled (2011), of which only Botswana and South Africa from sub-Saharan Africa
TIMSS	Mathematics, Science	Grades 4, 8, advanced[1]	Every 4 years	Grade 4: 36 countries (2007) Grade 8: 48 countries (2007) Grade 12: 10 countries (2008) mainly industrial
PISA	Reading, Mathematics, Science	15-year olds	Every 3 years	65 countries (2009) 30 OECD countries; 36 other

[1] *Final year of secondary school/first year of tertiary education.*

When students in low-income countries have participated in global international assessments based on achievement standards of industrial countries, they have performed poorly. For example, in TIMSS 2007, no students from **Botswana, Ghana, Morocco, Tunisia, or Yemen** reached the advanced international benchmarks in Grade 8 mathematics or science. Less than 4% of students even reached the 'high' benchmark.

Regional assessments: *Options*

Three *regional* assessments address the issue of the inappropriateness of global assessments for many countries by confining participation to countries in the same general region that are similar in their culture and economic development.

- *Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ)* is a grouping of 15 ministries of education in Anglophone countries in southern and eastern Africa that work in collaboration with the International Institute for Educational Planning (IIEP) in Paris. Although individual countries have produced a national report which does not include any comparative data, its website contains statistical and mapping tools for the regional comparisons, and SACMEQ is generally regarded as a regional assessment. The main focus of the consortium is on capacity development in policy makers and the administrators of the assessment in member countries.

- *Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN (PASAC)* is conducted under the auspices of the Conférence des Ministres de l'Education des Pays ayant le Français en Partage. Francophone countries across Africa participate. Pupils are assessed at the beginning and end of the year to assess growth in achievement. Individual countries have produced national reports which contain some comparative data. (See overview report.)

- *Laboratorio Latinoamericano de Evaluatión de la Calidad de la Educatión (LLECE)* is a network of national systems of education in Latin America and the Caribbean which is co-ordinated by the UNESCO Regional Office for Latin America and the Caribbean. LLECE conducted the first and second study, Segundo Estudio Regional Comparativo y Explicativo (SERCE). LLECE is both an assessment network and an assessment system intended to evaluate the quality of K-12 education in Latin America. It is now conducted at five yearly intervals, with grades 3 and 6, in Spanish, Portuguese and mathematics.

Table 5. Regional Studies of Student Achievement

	Construct/ Curriculum Area Assessed	Age/ Grade	Frequency	Participating Countries
SACMEQ	Literacy, Numeracy	Grade 6	1995-1998 2000-2002 2007	15 countries in southern and eastern Africa (2007)[1]
PASEC	French, Mathematics (+ national languages)	Grades 2 and 5	Annual (not in all countries)	21 countries in Francophone Africa
LLECE	Language (Spanish, Portuguese), Mathematics	Grades 3 and 6	Every 5 years[2]	16 countries (and one state) in Latin America and the Caribbean

[1] *7 countries in 1995-1998 and 14 countries in 2000-2002;*
[2] *The first LLECE study was conducted in 1997 at grades 3 and 4; LLECE's second assessment (SERCE) was conducted in 2007. It has been decided that from then on, assessments will be conducted every five years.*

Global and regional assessments: *Decisions*

Consideration of participation in an international assessment should be informed by answers to the following questions:

- Are policy/decision makers satisfied that the **achievements of students** (including the full range of these achievements) **will be adequately represented** in the assessment instruments? Studies that have compared the achievement domains assessed in TIMSS with the achievement domains of a national assessment or with national curricula have revealed considerable divergences. Thus, data from an international study may not provide a good indication of what students in a particular education system have learned, and national-level analyses based on the data may not be appropriate.
- Is there provision for representatives of one's education system to **participate in the design** of the assessment (eg, to contribute to test development; to try out items before selection for the final form of the test)?
- Is it envisaged that participation in an international assessment will contribute to the **development of local capacity** (e.g., in test development, sampling, statistical analysis)? Has existing capacity within the country (including capacity in agencies/institutions outside the ministry of education) been assessed? Would this represent the best way of achieving the goal? Will external technical assistance be required over the short or long term?
- Are policy/decision makers satisfied that it will be possible to **meet international standards** (e.g., in sampling, translation of instruments)?
- Are policy/decision makers satisfied that **meeting deadlines** for an international assessment is feasible in light of possible shortage of administrative personnel and poor communications infrastructure?
- Has the overall **cost of participation** in an international assessment been **compared** with the cost of a national assessment?

If an international assessment is being carried out in addition to a national assessment, what **additional information** is the international study expected to provide (eg, comparisons with performance in other education systems)?

8. What administrative and technical decisions apply?

Following a decision to carry out a national (or international) assessment, it is desirable that the ministry of education appoint a national **steering committee** to oversee its implementation. In addition to representatives of the ministry, one would expect to see on the committee, representatives of teacher unions, teachers, teacher trainers, curriculum personnel, and school managers.

A further decision involves identifying an **agency to carry out the assessment**. Countries have adopted different practices, assigning responsibility to a variety of organisations/groups: a team set up within the ministry, a unit set up in an examinations board, a university, a research centre, a consortium of groups, a national team supported by some international technical assistance, and non-national technical teams.

Skills to carry out the following tasks will be required:

- design and construction of assessment instruments;
- design and construction of background questionnaires;
- probability sampling;
- organisation of testing in schools;
- data preparation;
- data analysis;
- report writing;
- Dissemination.

Some of the skills and resources listed above will be present in, for example, a national examinations board. However, the design and implementation of a national assessment requires specialist skills which may not be sufficiently developed locally. For example, the **quality of data** from the assessment will depend on the quality of the sample. Drawing a **representative sample** which will give results of known precision is a highly technical task. If the sample is not drawn properly then the results will be of dubious value. This is often an area where enlisting external **technical assistance at the earliest opportunity** may be of great value.

In addition to sampling, a national assessment also requires special approaches to **data analysis**. In particular, classical test theory (CTT) and the item characteristics commonly used by examination boards have severe limitations when it comes to producing **meaningful scales of achievement and/or linking different tests**. For these purposes, international and national assessments commonly use item response theory (IRT). This is a sophisticated mathematical model which allows student ability to be placed on a meaningful scale which is directly related to the difficulty of the items on the test. This, too, is an area where enlisting external technical assistance may prove invaluable in both conducting surveys and building local capacity.

The **administrative and logistical difficulties** in conducting an assessment should not be underestimated. For example, considerable space will be required to prepare materials for schools and to process them after administration. This space is most likely to be found in an institution that carries out large-scale testing in schools (eg, a research centre, an examinations board). The timing of a national assessment may have to be decided in light of the demand for space and other facilities for examinations or other research studies. It is therefore necessary to coordinate the national assessment exercise carefully with respect to other demands in the academic calendar.

9. What is the cost?

National assessments costs

It has proved impossible to find comprehensive, reliable data on the costs of introducing and running a national assessment in developing countries. It seems that all too often no proper budgetary planning is done, and accounting records are incomplete. *Ad hoc* approaches are used to find money and many of the costs are hidden through the use of, for example, unpaid contributions by regional authorities, schools and teachers.

Assessing the cost of a national assessment is complex since no established formula exists. What is required is the collaboration of assessment specialists, policy makers, educators and economists. It is an iterative process in which specific inputs are modified until costs fit roughly within the budgeted amount.

Costs vary very much from country to country: this is a field in which averaging costs is not very informative. A more useful approach is to identify a specific case which matches the country of interest reasonably closely in terms of population distribution, size and assessment approach. The only published data on national, regional and international assessment costs currently available is that of Wagner 2010 (in press), summarised in Matrix 1. Of the countries and assessments reported, no more than 0.33% of the secondary education budget was spent on the exercise, and typically less than 0.1% of it, although there are gaps in the data and some costs cannot be captured. The absolute expenditure reported ranges from $122,190 to $2,768,571.

Data collection (Stage 4 in Table 6) is usually the most expensive component of an assessment. It is estimated that in the U.S. national assessment it absorbed 30% of budget. In some developing countries, it accounted for 50%. Assessments are more expensive in large countries than in small compact countries because of data collection costs. Salary levels and costs of services will vary with national economic conditions. If an assessment is carried out by a government-funded institution, there is likely to be a larger proportion of 'hidden' costs than if carried out in a non-government funded institution because, for example, salaries of some personnel will be paid from other sources.

Matrix 1: National, regional and international assessment costs (selected countries)

	National assessments			Regional assessments			International assessments					EGRA assessments [12]	
	SIMCE 2004 [1]	Honduras 2004 [2]	Uruguay 2003 [3]	PASEC 2010 [4]	SACMEQ III Swaziland 2007 [5]	SACMEQ III Tanzania 2007 [6]	PISA Chile 2009 [7]	PISA Mexico 2009 [8]	PISA Panama 2009 [9]	PISA Peru 2009 [10]	PISA Uruguay 2003 [11]	EGRA Liberia 2008	EGRA Nicaragua 2008
Test preparation	**258,236**	**174,275**	**21,528**	**34,164**	**12,561**	**12,666**	**26,448**	**100,301**	**61,475**	**47,956**	**12,357**	**29,345**	**10,882**
Creation and editing of test items	184,515			7,895		1,000	26,448	3,802	13,661				
Pilot testing	73,721			15,749	12,561	11,666		96,499	47,814			16,031	4,756
Training				10,520								13,314	6,126
Test application	**1,163,764**	**435,717**	**57,289**	**91,705**	**170,732**	**89,900**	**597,958**	**891,501**	**187,157**	**212,486**	**29,707**	**82,260**	**68,683**
Test design and editing	29,403			7,415		2,000	8,976		13,661	2,590		8,800	
Test printing	324,712			9,744	15,488	12,000		254,899	54,644	7,196		5,600	1,395
Printing of other materials	236,076				3,049	4,200		116,156	6,831				
Distribution to examiners	103,124			6,455	73,171	2,000		123,845	6,831				
Field testing	406,103			68,091	79,024	56,700	462,705	394,235	98,359	198,261		67,860	67,288
Control and supervision	64,346					13,000	126,277	2,366	6,831	4,439			
Processing and analysis	**382,239**	**130,721**	**26,272**	**12,624**	**454**	**33,300**		**167,782**	**128,414**		**22,838**	**13,533**	**5,734**
Coding and digital input	216,048			12,624		33,300		56,899	114,753			13,533	5,734
Marking open-ended questions	166,191				454			110,883	13,661				
Additional analyses													
Dissemination	**100,567**	**130,721**	**531**	**32,193**	**4,195**	**2,000**	**49,912**		**34,153**	**3,865**	**14,092**	**1,850**	
School communication	100,567				4,195	2,000			34,153	3,865		1,500	
Report production and distribution							49,912					350	
Public relations retainer													
Subtotal	**1,904,806**	**871,434**	**105,620**	**170,686**	**187,942**	**137,866**	**674,318**	**1,159,584**	**411,199**	**264,307**	**78,994**	**126,988**	**85,299**
Institutional costs	**938,766**			**12,481**	**24,878**	**25,500**	**179,233**	**490,203**	**94,261**	**20,473**		**103,520**	**87,157**
Personnel - in project budget	796,864			2,737	17,561	10,000	179,233	321,246	73,769	9,324		101,858	83,675
Personnel - contributed								107,286		11,149		1,403	2,500
Infrastructure - in project budget	35,369					5,000		2,743	6,831				
Infrastructure - contributed													
Equipment - in project budget	106,533			9,744	7,317	10,500		58,928	13,661			259	982
Equipment - contributed													
Test fees							49,863	118,599			43,197		
Other	20,028			2,043			72,494		13,661	2,000		10,619	6,958
TOTAL	**2,863,600**	**871,434**	**105,620**	**185,210**	**212,820**	**163,366**	**975,908**	**1,768,386**	**519,121**	**286,780**	**122,191**	**241,127**	**179,414**
Total Students	300,000	45,657	12,993	5,400	4,155	3,000 [13]	5,700 [14]	45,079	42,000	7,967	5,797	3,770	5,760
Total Schools												240	120
Cost per student	9	50	8	34	51	55	171	40	12	36	21	34	15
Cost of educating a student	767	130	484		66			9,439	1,023	396	479		
Cost of testing as % of total budget for one grade	0.83	2.63							1.20838				
Cost of testing as % of total secondary education budget	0.17	0.33	0.07					0.001767	0.04419		0.08		

[1] Source: Wolff 2007, p. 6 (for 2004 SIMCE test). Original figures for all national assessment data above (namely SIMCE 2004, Honduras 2004 and Uruguay 2003) and PISA Uruguay 2003 were published in Wolff 2007 in local currencies. In order to facilitate comparison across assessments in this table, Wolff's figures were converted from the original currency to the average annual market rate of USD. For the SIMCE 2004, Wolff used 2002 figures, in Chilean Pesos. Thus, these 2002 peso figures were all converted to the 2002 average annual market rate for USD (677.4916667 Chilean Peso to 1 USD). See also footnotes 2, 3 and 5.

[2] Source: Wolff, 2007, p. 13; 2004 17.68 Honduran Lempira to 1 USD.

[3] Source: Wolff, 2007, p.11; 2003 28.24279 Uruguayan Peso to 1 USD.

[4] Source: PASEC 2010 technical report (received via personal communuication, P. Varly, May 2009). Converted from Euros to USD, 2009 annual rate.

[5] Source: Personal communication, A. Mrutu, August 2009.

[6] Source: Personal communication, J. Shabalala, August 2009.

[7] Source: Personal communication, E. Lagos, September and October 2009.

[8] Source: Personal communication, M. A. Diaz, September 2009.

[9] Source: Personal communication, Z. Castillo, September 2009.

[10] Source: Personal communication, L. Molina, September 2009.

[11] Source: Wolff, 2007, p. 14; 28.24279 Uruguayan Peso to 1 USD (2003).

[12] Source: Personal communication, A. Gove, August 2009.

[13] Estimate, based on SACMEQ II sample of 2854.

[14] Estimate, based on email of E. Lagos, October 2009.

Source (personal communication): Wagner, D. A. (2011, in press). *Smaller, quicker, cheaper: Improving learning assessments for developing countries*. Paris/Washington: IIEP-UNESCO/Fast Track Initiative.

See also: Wagner, Daniel A. (2010) *Quality of education, comparability, and assessment choice in developing countries.* COMPARE: A Journal of Comparative and International Education, Vol 40, No. 6, 741 - 760.

Table 6 represents two extremes in a hypothetical national assessment. One is contracted out, and involves a complex test administered to all students (census), using state of the art methodology and technology. The other is administered by a government agency and is smaller in scope and ambition. Differences between the two are identified for eight major parameters of an assessment. On the basis of contextual factors, objectives and constraints (human, financial, material), an intermediate position is usually settled on.

If a country is unable to establish a budget line with sufficient funds to implement and sustain an on-going programme of national assessments, it should probably not embark on the exercise.

Table 6. Options and Cost Drivers at 8 Stages of a National Assessment

Options/key parameters	Higher specifications and/or higher cost drivers	Lower specifications and/or lower cost drivers
Stage 1. Governance ▸ Implementing agency capacity		
Public vs private implementing agency	*Non-government funded institution* Subcontracting of most of the work to external agencies and international/ local consultants. Accounting for inputs on timesheet basis can drive up cost.	*Government-funded institutions* Direct management of all tasks and operations at a national centre by civil servants. Test administration is part of the normal duty/routine activities of officials identified to carry out the study. Some costs absorbed/hidden.
Human resources	Large professional and administrative team contracted and mobilised.	Leadership undertaken by a single academic, official or a small coterie, plus research assistants.
Material resources	Inadequate ICT resources (telephone, computers, printing machines, etc) will increase costs.	Effective computerisation reduces costs.
	Implementing agency peripherally located or with difficult transport links across the country.	Implementing agency centrally located or with good transport links across the country.
Stage 2. Sampling ▸ Target population		
Sample size	Census-based assessment.	Sample-based assessment
Sample stratification	Robust sub-national data for various types of population (public/private; urban/rural; majority/minority; dominant language/ alternative language).	National-level data only.
	Targetting an age level.	Targetting a grade level.
Stage 3. Conceptualising ▸ Development of survey instruments		
Content	Testing several curriculum areas/constructs. Extensive number of curriculum domains assessed.	Testing one curriculum area/construct. Limited number of curriculum domains assessed
Construction	Competency-based test	Content-referenced test
Translation	Tests, questionnaires and manuals translated	All materials in one language

Options/key parameters	Higher specifications and/or higher cost drivers	Lower specifications and/or lower cost drivers
Stage 4. Administration ▸ Data collection		
Sample size and distribution	Large country/low density of population; long distance to travel and difficulty in gaining access to schools; large numbers of small schools.	Small/compact country.
Human resources	High number of data collectors.	Teachers administer the test, although consider the cost of lost teaching time.
Stage 5. Processing ▸ Data entry/cleaning		
Sample size	Population, or large sample.	Small sample.
Characteristics of tests	Open-ended test.	Multiple-choice test.
Scoring technology	Manual scoring in high wage-labour contexts. Note: machine scoring is a false economy unless large economies of scale can be reaped.	Machine scoring (assuming technology costs relatively low/ labour costs relatively high, and the capacity to operate and maintain the investment in technology exists).
Quality assurance	Stringent quality assurance mechanisms (data and marker checking).	Minimal quality assurance.
Stage 6. Analysis ▸ Identify significant results		
Human resources capacity	Contracting out/qualified data analysts.	In-house staff used.
Characteristics of tests	Complex test (multiple test booklets).	Simple test.
Technology available	Use Item Response Modelling, Hierarchical Linear Modelling.	Basic statistical analyses.
Stage 7. Informing ▸ Reporting and Dissemination		
Number of adapted versions and translations	Several versions of the main report for targetted audiences (policy makers, teachers, general public, etc.)	Limited number of reports (main report, executive summary)
Communication strategy	Translation of reports and associated communication tools.	No translation.
Stage 8. Use of data and action ▸ Follow-up activities		
Secondary analyses of data	Making data available and encouraging/ commissioning secondary analyses.	Data stored but not explored further.
Action: Human/ material investment to address shortages of resources identified	Meetings, conferences to discuss results; training programmes to transfer/upgrade skills in key professional areas related to national assessment.	Limited follow-up activities.
Action: Education system reforms	Policy initiatives flow from key findings where applicable, eg diagnostic results on teaching practices (increased investment cost but also greater value for money from assessment).	Assessment not linked to policy. (Low investment cost but also low value for money from assessment).

Data on the cost of conducting national assessments is not generally available. The following examples from Nigeria are based on figures from a DFID-funded programme.

A. Monitoring of Learning Achievement:

Coverage: six states of Nigeria. 390 schools in total (various combinations of public and private schools by state, according to location.)

Sample: 40 pupils per school, 15,600 children in theoretical sample (somewhat fewer in achieved sample); each pupil completing two papers (maths and English) = 31,200 marksheets.

Method: Orally administered—highly intensive in personnel.

Total out-of-pocket expenses for administration:

approximately USD205,000.

Technical assistance: approximately USD135,000.

Total cost: approximately USD340,000 = approximately USD22 / pupil.

B. Teacher development needs analysis:

Coverage: five states of Nigeria, of which four were conducted on a probability sample basis, and one by surveying the entire primary and junior secondary teacher population of the state (19,125). The following figures refer only to the four sample-based states.

Sample: 9,200 teachers in four states.

Method: three hour, paper-based, authentic, ethically-sound assessment task (marking a grade 4 pupil's answer papers in maths and English; assimilating a variety of source material to plan a geography lesson; application of numeracy skills to simple performance statistics of students in a class.)

Total out-of-pocket expenses for administration: approximately USD200,000 (of which 50% borne by state governments).

Technical assistance: approximately USD85,000.

Total cost: approximately USD285,000 = approximately USD31 / teacher.

The only other specific information about the costs of national curriculum tests identified, is Whetton's figure of GBP40m (USD63.6m) to conduct the SATs (standard assessment tasks) in England, which have been administered annually to around 1.8m children at 25,000 schools, in several subjects and three age groups. This equates to approximately USD35 per child.

International assessment costs

In global international studies (prePIRLS, PIRLS, PISA, TIMSS), each participating country is required to cover its own costs of the study at the national level, and to contribute to the cost of co-ordinating the study at an international level. PISA fees, which are calculated according to the size of a country's economy, range from Euro 50,000 to 600,000 per year (at the time of writing, Euro 1 = USD 1.22). By contrast, fees for prePIRLS, PIRLS and TIMSS are fixed for all countries. A participating country is expected to pay approximately USD 175,000 for prePIRLS or PIRLS, and USD 310,000 for TIMSS (2 grades). Compulsory attendance at international meetings adds to costs. IEA provides participating countries with a national budgeting framework and schedule, to assist with estimating the highly-variable in-country costs.

Fees to international bodies represent a significant share of the total cost of an assessment. In the case of **Peru** and **Uruguay**, payment to OECD for the management of PISA accounted for 21% and 35% respectively of total expenditure.

No direct fees are payable by countries participating in PASEC or SACMEQ. Latin and Caribbean countries have to contribute USD 10,000 per annum to participate in LLCE's regional assessment. However, this is not a significant proportion of the overall cost of the assessment. While country costs for these assessments are still relatively low, they have doubled since the 1990s.

10. How should the findings of national assessment be used?

Assessments provide information that can be used to **improve the quality of student learning.** Creating **awareness** of this is essential for optimal use to be made of the findings of a national assessment.

Assessments should not be seen as one-off, isolated exercises. To ensure that this is the case requires the assessment system to be **institutionalised and integrated** into the structures and processes of government policy formation, decision making, and channels of resource allocation.

This, in turn, will require investment in development of **institutional capacity to absorb and use the information** provided by the assessment. This includes the integration of assessment information into Educational Management Information Systems.

It is important that national assessments be **aligned** with other aspects of the education system, including other assessment systems, curricula, teacher education, school capacity building, and measures to address inequalities.

Ensuring that information is provided to all involved in policy and management in an **intelligible form** will require the preparation of a number of reports and briefing papers in addition to a main or technical report.

In this paper policy makers and education managers are regarded as the primary **audience for learning assessment findings**. This is not to ignore the fact that many other stakeholders in the education system also have an interest in the findings. Results should be presented and made available to them in an intelligible form. Of these, teachers are a key constituency since student learning is unlikely to improve unless policies and strategies are developed to change school and classroom practice. Special reports for teachers, reinforced by meetings in which findings are presented and discussed, are perhaps the most direct and effective way of bringing about desirable changes in the teacher-learner relationship.

In **the Gambia**, national assessments have been carried out in English and mathematics (grades 3 and 5) and to assess early reading skills using the Early Grade Reading Assessment protocol (grades 1, 2, and 3). Arising out of the poor performance of students, a task force consisting of senior officials was set up by government. Review workshops were held to identify gaps in instructional materials and teacher training curricula. Supplementary readers were produced, and a handbook on teaching early reading was prepared for use in teacher training courses.

In some countries, policy makers have expressed interest in using the results of national assessments to make **schools and, in some cases, individual teachers accountable for learning outcomes**. In order to do this, all students would need to be assessed (census) since in a sample-based approach most schools and classrooms would not be included. Secondly, making only schools accountable ignores the fact that the factors determining learner achievement are many, varied and complex. Thirdly, it is neither logical nor ethical to introduce an accountability framework unless teachers and head teachers are given meaningful control over decisions and resources affecting the quality of learning outcomes in their own classrooms and schools. Fourthly, the assessment used must meet exacting technical standards and be fairly administered in all locations if allocation of resources to schools and teachers' careers depend on the outcome. A valid accountability system needs to take all these factors into account be fore assigning credit or blame.

This is a highly technical exercise and where it has been tried (eg, in the contextual value-added systems in England and the USA) it has evoked much criticism. The **challenges of applying accountability frameworks** are multiplied in resource-constrainted education systems. Rather than be used initially as the basis for an accountability-based education system, national assessment could play a significant role in stimulating debate in civil society and government about clearly stated education standards and the need for evidence of learning outcomes. Such debate can reveal the links between (a) management of resources at the school level, (b) the extent of local-level responsibility for school performance, and (c) the need for accountability to follow empowerment of communities and head teachers, along with other supporting inputs. Community and NGO assessments (eg Uwezo, ASER), which do not conform to national assessment criteria but which should be well-calibrated to local conditions, can play a complementary role in this debate. It also points towards emphasising the responsibility of the state as ultimate guarantor of educational standards and quality for all children, rather than necessarily being the universal supplier of schooling.

11. Conclusion: guidance on options

Carrying out an assessment will always require a **compromise** between the ideal and what is possible. Achieving the well-defined objectives of even well-resourced international studies requires working within constraints of budget, time, national politics and ethical concerns. Conducting a national assessment of student achievements, especially in developing countries, involves making choices among the variety of approaches that exist, working with what is possible and available, and recognising the limitations implicit in the choices that are made.

The decisions made about the purposes and nature of any national assessment, or about a country's involvement in an international study, will depend on the particular information needs and priorities of policy/decision makers, as well as the circumstances in which they operate. However, when a country is at an early stage of developing a learning assessment system and has limited resources, certain options are generally preferable. The following table lists preferences, all other things being equal, among the choices that are available, and which meet the objective of informing policy and practice regarding the quality of student learning in the education system.

Table 7. Preferences regarding national and international assessment choices

Preference	Reason
Carry out a national rather than an international assessment, unless the international assessment is particularly well suited to the country's needs.	More likely that national needs and concerns will be addressed, that the assessment framework will reflect the full range of student achievements in the education system, that relevant background data will be collected, and that local stakeholders are involved in all aspects of the process.
Involve policy/decision makers, in collaboration with other stakeholders, in specifying the issues they expect an assessment to address. Maintain the involvement of a steering committee throughout the assessment process.	Focuses the assessment on the issues that are of concern to policy/decision makers and other stakeholders, increasing the likelihood that findings will be acted on.
Accord priority to assessing literacy and numeracy.	Recognises the crucial importance of literacy and numeracy skills for learning in the education system.
Carry out an assessment at an early stage in the education system (eg, about two years after pupils have begun to read in the language of the assessment).	Facilitates the detection of deficiencies that might underlie later difficulties, at an early stage when remedial action can be taken.
Use a sample of achievement that is sufficiently wide to provide diagnostic information on students' performance.	Provides diagnostic information on students' performance that can be used to inform the practice of teaching, curriculum development, and preservice and inservice teacher education
Report achievement for constituent domains and/or in terms of proficiency levels, rather than as simple average scores.	Provides a basis for policy and remedial action, as it can identify what precisely students know and can do, and what they do not know and cannot do.
Base the assessment on a sample rather than on a census.	Less expensive and more efficient.
Develop assessment capacity (in policy formation and technical skills) at a national and/or regional level.	Can be tailored to take into account the availability of local capacity and the areas in which further development is needed.
Carry out an assessment every three or four years, unless circumstances dictate otherwise.	Should be sufficient, as achievement levels generally change slowly.
Prepare a number of reports, in addition to the technical report.	To meet the specific and varied needs of stakeholders: policy/decision makers, teachers, teacher trainers, curriculum developers, the general public.
Make data from an assessment available for further analysis to universities, research organisations, and regional education authorities.	To exploit the rich data which an assessment can provide and which because of time, technical and financial constraints, the body that carries out the assessment may not be able to do.

Annex 1. NAEP Mathematics Achievement Levels, Grade 4: United States

Level	Expected achievement at grade 4
Basic	Students should show some evidence of understanding mathematics concepts and procedures in the five NAEP content areas. They should be able to estimate and use basic facts to perform simple computations with whole numbers, show some understanding of fractions and decimals, and solve some simple real-world problems in all NAEP content areas. They should be able to use, although not always accurately, four-function calculators and rulers. Their written responses will often be minimal and presented without supporting information.
Proficient	Students should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content areas. They should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators and rulers appropriately. They should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented with both supporting information and explanations of how the solutions were achieved.
Advanced	Students should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP content areas. They should display mastery in the use of four-function calculators and rulers. They are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, the solutions were achieved. They should go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.

Source: U.S. National Center for Education Statistics 2006a.

Annex 2. Reading Skill Levels in National Assessment, Grade 5: Vietnam

Skill level	Achievement	Percent of students at this level	Standard error
1	Student matches text at word or sentence level, aided by pictures. Skill is restricted to a limited range of vocabulary linked to pictures.	**4.6**	**0.17**
2	Student locates text expressed in short, repetitive sentences and can deal with text unaided by pictures. Type of text is limited to short sentences and phrases with repetitive patterns.	**14.4**	**0.28**
3	Student reads and understands longer passages. Student can search backward or forward through text for information and understands paraphrasing. An expanding vocabulary enables understanding of sentences with some complex structure.	**23.1**	**0.34**
4	Student links information from different parts of the text. Student selects and connects text to derive and infer different possible meanings.	**20.2**	**0.27**
5	Student links inferences and identifies an author's intention from information stated in different ways, in different text types and in documents where the message is not explicit.	**24.5**	**0.39**
6	Student combines text with outside knowledge and hidden meanings. Student identifies an author's purposes, attitudes, values, beliefs, motives, unstated assumptions, and arguments.	**13.1**	**0.41**

Source: World Bank 2004, vol. 2: table 2.1. Sums to 99.9% because of rounding errors.

Annex 3. Glossary

Classical Test Theory (CTT) A measurement theory which consists of a set of assumptions about the relationships between actual or observed test scores and the factors that affect the scores. It is used for measuring and managing test and item performance data. In contrast to item response theory, it comprises a set of more traditional psychometric methods.

Construct A hypothesised ability or mental trait that is used to explain performance on an assessment. Constructs cannot necessarily be directly observed or measured. In a language assessment, in addition to language ability itself, motivation, attitude and acculturation are all relevant constructs.

Cut-score A selected point on the score scale of a test, such that scores at or above a particular point are interpreted differently from scores below that point. In achievement tests, sometimes there is only one cut-score dividing the range of possible scores into two regions (eg, 'passing' or 'failing', 'mastery' or 'non-mastery'). Sometimes more cut-scores are used to define categories and establish performance standards (eg, 'advanced', 'adequate', 'basic', 'inadequate'). Cut-scores should be based on a generally accepted methodology and reflect the judgments of qualified people.

Hierarchical Linear Modelling Also known as multi-level analysis, is useful for understanding relationships in hierarchical data structures (e.g., students nested within classrooms, classrooms nested within schools, schools nested within districts). It is an advanced form of regression, allowing variance in outcome variables to be analysed at multiple hierarchical levels, whereas in simple and multiple linear regressions all effects are modelled to occur at a single level.

Item Response Theory (IRT) A group of mathematical models for relating the performance on a test item to the test taker's level of performance on a scale of the ability or trait being measured. An advantage of IRT over more traditional types of analysis is that it has the potential to provide item characteristics that are independent of the candidates who took the tests. By modelling the response of a test taker of given ability to each item in the test, an item has a known probability of being correctly answered by an individual of a given ability level.

Item characteristics The type and functioning of a single task or question within a test instrument, and the assumptions underlying its use, including its difficulty, level of discrimination, and degree of bias. An individual's probability of success on an item is said to be governed jointly by his/her ability and by the difficulty and discrimination of the item. An appropriate achievement test is likely to consist of items with a variety of difficulty levels and effective discriminations.

Probability sample A sample that is selected in such a way that every element of the population has a known probability of being included.

Standardised A test procedure in which the questions, papers, administration conditions, scoring and interpretation of results are applied in a consistent and pre-determined manner for all test-takers. Interpretation of standardised test scores can be norm-referenced or criterion-referenced. Norm-referenced standardised tests allow for comparisons of results between students which cannot reliably be inferred from non-standardised tests. Criterion-referenced standardised tests allow for identification of students who have attained a cut-score with respect to the skill or curriculum area being tested (the criterion), irrespective of the performance of their peers. Standardised scores allow for placement of students on a readily-understandable scale, commonly centred at 100 to represent the average nationally standardised score for the population concerned. Such comparisons cannot meaningfully be made using raw scores or percentage scores.

Annex 4. Selected Readings

The following are relevant to several sections in the Note.

C. Bangay. (n.d.). Assessment – *Ways and whys*. London: Department for International Development.

V. Greaney & T. Kellaghan (1996). *Monitoring the learning outcomes of education systems*. Washington D.C.: World Bank.

V. Greaney & T. Kellaghan (2008). *Assessing national achievement levels in education*. Washington D.C.: World Bank.

P. Ravela, P Arregui et al. (2008). *The educational assessments that Latin America needs*. PREAL Working Paper No. 40.

The following have particular relevance to specific sections in the Note.

Section 2

R. Coe (1999). *Changes in examination grades over time: Is the same worth less?* Curriculum, Evaluation and Measurement Centre, University of Durham.

T. Kellaghan & V. Greaney (2004). *Assessing student learning in Africa*. Washington, D.C.: World Bank.

Section 3

J.E. Cohen, D.E. Bloom & M.B. Malin (2007). *Educating all children: A global agenda*. Cambridge MA: MIT Press.

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

N.S. Raju et al (eds.) (2000). *Grading the Nation's Report Card: Research from the evaluation of NAEP*. Committee on the Evaluation of National and State Assessments of Educational Progress, National Research Council. The National Academies Press.

Section 4

H. Braun, A. Kanjee, E. Bettinger & M. Kremer (2006). *Improving education through assessment, innovation, and evaluation*. Cambridge MA: American Academy of Arts and Science.

D. J. Johnson (2008). *An Assessment Of The Development Needs Of Teachers In Nigeria*, Kwara State, Final Report, CUBE.

Section 5

B. Alvarez & M. Ruiz-Casares (Eds) (1998). *Evaluation and educational reform*. Washington D.C.: Academy for Educational Development.

P. Anderson & E. Morgan (2008). *Developing tests and questionnaires for a national assessment of educational achievement*. Washington D.C: World Bank.

H. Braun (2004). *Reconsidering the impact of high-stakes testing*. Education Policy Analysis Archives 12(1).

L. Connors, D. Putwain, K. Woods & L. Nicholson (2009). *Causes and consequences of test anxiety in Key Stage 2 pupils: The mediational role of emotional resilience*.

T.N. Postlethwaite & T. Kellaghan (2008). *National assessments of educational achievement*. Brussels: International Academy of Education; Paris: International Institute for Educational Planning.

RTI International (2009). *Early grade reading assessment toolkit*. Washington D.C., World Bank.

Section 6

T. Kellaghan, V. Greaney & T.S. Murray (2009). *Using the results of a national assessment of educational achievement*. Washington D.C.: World Bank.

Section 7

N. Altinok (2008). *An international perspective on trends in the quality of learning achievement (1965 -2007)*. Paper commissioned for the EFA Global Monitoring Report 2009, Overcoming Inequality: Why governance matters.

A.E. Beaton, T.N. Postlethwaite, N. Ross, D. Spearitt & R.M. Wolf (1999). *The benefits and limitations of international educational achievement studies*. Paris: UNESCO: International Institute for Educational Planning.

W.B. Elley (2005). *How TIMSS-R contributed to education in eighteen countries*. Prospects, 35, 199-212.

Section 9

L. Ilon (1996). *Considerations for costing national assessments*. In P. Murphy et al (Eds), National assessment: Testing the system. Washington D.C.: World Bank.

M.T. Siniscalo (2006). What are the national costs for a cross-national study. In K.N. Ross & I.J. Genevois (Eds), *Cross-national studies of the quality of education: Planning their design and managing their impact* (pp. 185-209). Paris: UNESCO: International Institute for Educational Planning.

L. Wolff (2007). *The costs of student assessment in Latin America. WP no 38*. Washington D.C.: Partnership for Educational Revitalization in the Americas (PREAL).

Section 10

P. Arregui & C. McLauchlan (2005). *Utilization of large-scale assessment results in Latin America*. Report prepared for the Partnership for Educational Revitalization in the Americas (PREAL) and the World Bank Institute, Washington D.C.

T. Kellaghan, V. Greaney & T.S. Murray (2009). *Using the results of a national assessment of educational achievement*. Washington D.C.: World Bank.

Wagner, D. A. (2010). *Smaller, quicker, cheaper: Improving learning assessments for developing countries*. Draft. Paper commissioned by the IIEP-UNESCO and the Fast Track Initiative. Philadelphia: International Literacy Institute.

Whetton, Chris (2009). '*A brief history of a testing time: national curriculum assessment in England 1989-2008*', Educational Research, 51: 2, 137 — 159.

About this paper

This paper was written by Thomas Kellaghan, George Bethell and Jake Ross, with contributions from Colin Bangay, Elsa Duret, Jenny Hsieh, Daniel Wagner and Christine Wallace.