MODULE 6

Accountability

Overview

Most people, lay and professional alike, agree that the education sector should receive a large percentage of a country's investment. Since the advent of human capital theories this has indeed been the case in many countries of the developing world. However, emphasis on such heavy capital outlay must at some point be shown to be justified, and there must be evidence that money has been well invested, because for most of these countries resources are scarce. Yet many countries have so far been unable to undertake a thorough evaluation of their education systems. Such an evaluation would seek to determine whether the system had been *efficient* in its use of the country's resources, whether it had made *equity* gains for different sectors of the population and whether it had been used to improve the *quality* of education in the country. These are the accountability issues on which an education system needs to be evaluated.

The obvious task in evaluating an education system would be to assess its outcome – i.e. student achievement. How is this to be done? Clearly, testing is the answer. Thus, this chapter will discuss some issues relevant to the design and construction of tests. Tests, however, serve different functions. For example, they may be used for selection purposes or for the gaining of credentials, or they may be designed to give an indication of the health of the system. Some idea of the basic differences in tests will be developed. Once a test has been implemented and scored through, how are the results to be interpreted? This takes us into the highly contentious world of 'standards'. Here we find that many people have only a hazy idea of what standards are and yet continually make judgements based on them. This module will show that standards have to be carefully selected, assessed and monitored if they are to be used in evaluating the system.

If we succeed in all these efforts, the unhappy fact remains that we still do not have a complete assessment of the education system. We need to engage in cross-national comparisons to see how our efforts measure up with those of others striving for similar goals – otherwise, it would be only a self-contained endeavour. This chapter will look at the kind of information generated by surveys conducted by the IEA (International Evaluation of Educational Assessment) as a measure of how the system is functioning.

Objectives

When s/he has completed this module, the participant should be able to:

1. Justify, in the context of educational financing, the need to set up evaluations of education systems.
2. Explain how national attainment tests differ from those designed for other purposes.
3. Interpret indicators of system performance.
4. Critically appraise the notion of 'standards' in the assessment of educational achievement.
5. Justify the need to include comparative international educational indices of assessment for purposes of accountability.
6. Demonstrate an awareness of the procedures and mechanisms which guarantee accountability, not just for compliance, but in the interests of efficiency, equity and improvement in the education system.

Accountability and educational financing

The education system can be assessed on its efforts to realise efficiency, equity and quality or improvements, in the delivery of education to the country. These are the reasons for which it was financed. These are the tasks an education system is required to perform and for which it is accountable .

Efficiency has to do with whether the system is maintaining or increasing its output of graduates at every level in a cost-effective manner. It would thus examine student-teacher and enrolment ratios, different ways of transmitting curricula, different curricula, drop-out and repetition rates, and graduation rates. Such indicators need to be periodically monitored to give policy makers a real sense of whether the system is doing what it is supposed to do cost-effectively. They may well suggest that more funding may be necessary in some areas for the system to become more efficient in producing output.

Equity is concerned with whether all groups in the society are being treated fairly by the education system. Hence, no group should be over-represented as graduates at any level of the system or appear to be confined to certain types of schools. This hinges on the clear correlation between educational credentials and well-paying jobs, leading to a higher standard of living. Equity, then, is really a means of measuring to what extent all groups are able to access social mobility. Governments expect that the heavy financial

investment they have made in education is one indirect, long-term way of addressing this issue and the system must therefore deliver.

Quality indicators have to do with pedagogy and the nature of the education that is being transmitted. Are the curricular programmes designed for, and relevant to, the age and interests of the students? Is the teaching innovative and stimulating? Are most teachers trained? Is the learning and knowledge helpful in developing critical thinkers and productive citizens? The improvement of instructional practices also has implications for improving the efficiency of the system, so that efficiency and quality indicators are closely related.

All countries seem to invest heavily in education. National goals may vary from country to country, but all countries want their education systems to be efficient, equitable and capable of delivering a 'quality' education. Hence, the system has to be evaluated to see how well it is performing on these indicators. Especially in small developing countries, planners and policy makers in need to become aware of, and insist on the need for, the establishment of procedures and mechanisms to guarantee accountability in the education system.

Accountability and achievement indicators

The issue of accountability in education is bound up with the dominance of academic achievement as its major output indicator. The notion of achievement is not at issue because this is, after all, the major project of schools – to educate students and then test them on what they have learned. However, there is some dissatisfaction with the specific indicator most in use today, that of cognitive or intellectual achievement. Box 1 below discusses some of the issues surrounding output indicators.

Box 1

Choice of Outcome Indicators

To enumerate the outcomes of education about which it might be useful to have empirical information in terms of the many aims that have been posited for education would be an endless task. Aims frequently suggested include the development of literacy and numeracy skills, the development of aesthetic areas of experience, preparation for life in a democratic society, preparation for the world of work, development of character and moral sensitivity and personal self-fulfilment. Aims (and expected outcomes) may differ for different ages and students. Given the range of educational aims and the complexity and difficulty of measuring outcomes, some selection has to be made in deciding what outcomes should be measured for use in an indicator system. All we can hope for is information on a limited number of indicators rather than a description of all possible outcomes.

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

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

Source: Adapted from Greaney and Kellaghan, 1996, pp. 4–5

Cognitive testing, however, is not without its array of problematic issues. Since the evaluation of achievement on a national scale is likely to be an expensive undertaking, decisions have to be made about selecting some

areas of competence to test. Which subject areas are more inclined to give key insights into the performance of the system against the goals of efficiency, equity and quality? The key subject areas that are normally chosen are English, mathematics, science and social studies. They represent areas of the curriculum that are considered core subjects and therefore most students have been exposed to them. National testing in subjects offered only by some students, for example art or technical-vocational studies, is of limited value.

Important decisions also have to be made about the frequency with which this exercise should be undertaken. Is it feasible to run the same tests every year to get an idea of trends over time, or is it just as feasible to do so in cycles four or five years apart? What kind of designs should be used in the assessment of student achievement? Cross-sectional designs will garner information on how students are performing in certain subjects across the nation at one point in time. Longitudinal designs, on the other hand, provide data for the same (or different) cohorts at certain intervals over time? Which should be chosen if both analysis of outputs and monitoring of learning is required?

Summary

This introduction to testing in the context of establishing accountability of the educational system, shows it to be an extremely specialized subject. Competent personnel in the field of testing and measurement are needed to design procedures and mechanisms that facilitate meaningful evaluation of national education systems. The sections below will address some of the specialist knowledge needed in the field of test construction and information needed in the analysis of results. The latter, however, is not an 'exact science' as it rests on values inherent in the selection, monitoring and evaluation of standards. Hence, accountability in educational systems not only rests with statisticians and measurement personnel but also with those knowledgeable about the workings of the whole system – the planners and policy makers in education and finance.

Testing

Distinguishing between different kinds of tests – Box 2 below summarises the variety of tests that exists in a school system and the purposes for which they have been constructed. Tests designed for evaluation of national educational systems must differ in fundamental ways from tests taken by primary school students to select those to go on to secondary education, or tests taken to obtain credentials at the end of secondary school.

Box 2

Testing and Evaluations of Educational Systems

Tests are the most common measure of student learning and are typically used to monitor learning. Tests are defined as 'any series of questions or exercises or other means of measuring the skill, knowledge, intelligence, capacities or aptitudes of an individual or group' (Anderson, Ball and Murphy, 1975, p. 425). Education achievement tests focus on skills, knowledge and capacities that students acquire through education (schooling). They do not seek to measure intelligence or aptitude.

Psychometricians construct educational achievement tests for many different purposes, of which the most common are:

- Monitoring achievement trends over time;
- Evaluating specific educational programs or policies;
- Holding schools, regions, and other administrative sub-units accountable for student achievement;
- Selecting students for further education;
- Certifying student achievement;
- Diagnosing individual learning needs.

The purpose of the test determines its overall design, for example, the purpose of selection tests is to discriminate between individuals, and hence the test will include many items of a difficulty known to discriminate between those who should be selected and those who should not. As a consequence, the range of competencies tested will be highly constrained. By comparison, the purpose of a test designed to monitor national achievement is to assess a large number of competency areas, but to provide little information about specific individuals. It will include items covering many competencies, but its power to discriminate among individuals will be constrained. Tests designed to monitor trends over time will repeat at least some items, whereas tests designed only to select once need not be equated.

Source: Adapted from Lockheed, 1996, pp. 10–11

Activity 1

Consider the kinds of tests that occur in your country or region. Classify them according to the six criteria listed in Box 2. Is there evidence of any evaluation of the national education system? If there is not, what are the reasons for this? If there is, investigate:

1. *How recently this practice was established;*
2. *What subject areas have been tested;*
3. *How the actual test items differ from those in other kinds of tests;*
4. *At what levels in the system it has been considered feasible to test.*

National assessment

National assessment testing also differs from high-stakes selection or credentialling tests in that a great volume of back-up data is collected so that test results can be appropriately interpreted by planners and policy makers, educators, parents and the public in general. Box 3 below elaborates on the data required by national assessments and how they are used.

Box 3

National Assessments and Background Data

Educational assessments include not only measures of student learning, but also of student background and school inputs and processes that enable the assessment data to be used for policy analysis. Many researchers have sought to identify the minimum set of information that should be collected. Most recommend the collection of basic intellectual, socio-economic, and cultural background data on students. Examples include the students' sex, age, nutritional status, socio-economic status, language, attitudes, expectations and ability. Theory, research and experience suggest four areas about which data on classroom inputs should be gathered: the curriculum, instructional materials, adequate time for learning and effective teaching . . . Beyond these inputs are (a) managerial processes at the schools and within the classroom, such as tracking, school structure, and autonomy; and (b) school context conditions such as student access to knowledge, organisational press for achievement, and professional teaching conditions.

Source: Adapted from Lockheed, 1996, p. 14

Contextual information

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

Several kinds of contextual information are likely to be of value in providing clues to policy makers about the determinants of achievement, particularly determinants that might be alterable through changes in educational policy. It is interesting to know what students bring to school from their family and community backgrounds that may contribute to successful or poor school performance. One cannot assume, however, that factors found to be important in one cultural context will be of similar importance in other contexts. For example, the fairly consistent conclusions from industrial countries in Europe and North America that family size is negatively related to educational achievement has not been found in studies in Kenya or in Tanzania. In these countries a positive rather than a negative relationship was found between family size and educational achievement, measured in a variety of ways.

Source: Adapted from Greaney and Kellaghan, 1996, pp. 35–36

Activity 2

Investigate the research evidence on the distribution of educational resources (for example physical plant, curricula, libraries, laboratories) in different countries and their record of educational achievement.

- *Compare the data from developed and developing countries.*
- *Are there indications that a wide array of contextual data is valuable in formatting analyses of achievement patterns?*

Public examinations

Public examinations differ markedly from national assessments because their purposes vary. National assessments are conducted to monitor the levels of equity, efficiency and quality education in the system and require a host of contextual data. Public examinations discriminate between students who are then sorted and allocated within the system. They are regarded as high-stakes examinations and largely determine the life-chances of students. National assessments are low-stakes. They give only an indication about the functioning of the system, and only a sample of students need be tested. However, the tests are comprehensive in covering the curriculum because assessors need to know how well students are doing on all curricular objectives. Public examinations, on the other hand, involve all students in a year group and differentiate between students of different abilities rather than gauge whether curricular objectives have been met.

The nature of the instrument

Testing is a precise activity. The test is referred to as an instrument that measures performance or achievement. Boxes 4, 5 and 6 below discuss some of the criteria involved in the construction and administration of tests – i.e. standardisation procedures, reliability and validity considerations and item discrimination. These procedures are necessary to ensure that there is equity in testing – that all students sitting the examination are dealt with in a fair way. Such precautions are absolutely vital when tests are being administered across the country in all kinds of schools and contexts as well as when they are being administered over time to different groups so that comparability can be maintained.

Box 4

Testing and Standardisation Procedures

Tracking progress requires having tools capable of monitoring trends over time. Tests must be standardised with respect to content, behavioural objectives, format, administration procedures, and scoring. Standardising the content requires that the same or equivalent questions or performance tasks be posed for all students. Standardising test administration requires uniformity in the written and verbal instructions given to examinees, in the length of time afforded them, in the materials provided to them, and in the physical testing environment. Standardised scoring requires explicit, impartial procedures for correcting tests or judging performance.

Source: Adapted from Lockheed, 1996, p. 12

Activity 3

Unlike tests for certification and selection, test items for national assessments tend to be repeated. This is how standardisation of content is achieved.

- *Consider reasons why test security is not a problem in this instance?*
- *How are test items for certification and selection standardised for content from year to year?*
- *Investigate the range of different kinds of tests (for example, objective tests, essays, practical tests). Which of these lend themselves easily to standardisation of content, administration and scoring procedures?*

Other than being standardised, test instruments need to be valid and reliable, whether for public examinations or national assessments. The excerpt in Box 5 explains the difference between the two concepts, and the importance of both, in test construction.

Box 5

Testing and Considerations of Reliability and Validity

An instrument is reliable if it measures whatever it measures accurately and consistently, but it is valid if it measures what it is expected to measure and does so accurately and consistently. Consider the following example. Suppose a country's ministry of education has the following telephone numbers:

Minister	042-770-504, 042-771-345
Permanent Secretary	042-771-640
Secretary to the Minister	042-771-500

Suppose that when you dial the minister's number each time you reach the minister's secretary. The telephone has not given you what you wanted, but it has been accurate and consistent in what it has given. It is a reliable phone. However, it is not valid because it has not delivered what you wanted, namely, getting through on the minister's line. If, however, each time you dial 042-770-504 the phone on the minister's desk rings, then it is valid because it is giving you what you wanted in accurate and consistent manner. Observe also that, in this case, the

phone is also reliable because it is accurate and consistent. In general, if an instrument is valid, it is necessarily reliable. If, by contrast, when you dial 042-770-504 and reach 042-771-345, 042-771-640, 042-771-720, or 042-771-250 at various times, then the phone is neither reliable nor valid.

Reliability is usually established by using such methods as test-retest, equivalent forms, split half, and the Kuder-Richardson formula. As for validities, these may be of different types, for example, concurrent, predictive, face, and content. An important aspect of any national assessment is that the utmost care is taken to ensure the acceptability and accuracy of the outcomes.

Source: Adapted from Nwama, 1996, p. 25

Activity 4

- *Investigate at least one statistical method, mentioned in Box 5, of establishing test reliability. Create a test in a curriculum area in which you are familiar. Assess its reliability characteristics.*
- *Examine the differences between the different kinds of test validity mentioned above. Use the test you have created to illustrate at least one kind of validity check.*
- *Attempt to justify to an audience, who is unaware of the procedures involved in test creation, the importance of designing reliable and valid measuring tools.*

Whether high or low stakes, tests are instruments which must be precisely put together if they are to be taken seriously as generating informative output indicators. Standardisation procedures, issues of validity and reliability and item analysis then become major features of testing. Box 6 below discusses some of the concerns inherent in item writing.

Box 6

Testing and Item Analysis

Care is also taken to ensure that items are free from cultural, racial, religious, gender, and location bias. This is extremely important because research has shown that seemingly neutral achievement test items may actually favour one subcultural group or handicap another. In multi-ethnic and geographically diverse situations, as is the case in developing Africa, this factor does operate and test developers must recognise it and guard against it. This can be done through detailed item critiques and revisions.

Source: Adapted from Nwama, 1996, p. 26

Dangers of Cultural Bias in Testing

The following problem was given to a sample of adolescents and adults as part of a pretest in a study of basic learning needs in Bangladesh:

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

Succession and property rights under the laws of the two main religious groups that apply in Bangladesh would prompt respondents to offer answers other than 8 bighas. Under Islamic law, the mother gets one-eighth of the property and each daughter gets half as much as each son. Under Hindi law, the daughter receives nothing if there are sons in the family.

Source: Adapted from Greaney and Kellaghan, 1996, p. 60

Activity 5

Item writing is a technical and specialised task.

- *Research the topic of item writing and suggest other kinds of criteria (other than cultural bias) which are important in creating items for examinations.*
- *Review any test that you can procure from primary or secondary schools or from a public testing agency. Critically assess the test items against the criteria you have researched.*
- *Should informal classroom tests to monitor learning be designed with the same kind of rigour of formal tests? Survey some teachers in the system to get a sense of how they approach testing.*

Designs of national assessments

Countries may adopt a cross-sectional or longitudinal design, or some type of compromise in its evaluation of the system. Whatever design is adopted depends on what resources the country commands and what kinds of questions it wants answered. Although only a sample is involved in national assessments, considerable investment in the project is still required. Box 7 below explores some of the emphases, strengths and shortcomings of each type of design.

Box 7

Survey Design Issues

Implicit in a decision to do a cross-sectional study is a focus on the cumulative learning of children, either up to the particular grade of school or by a particular age. Implicit in a design to do a repeated-measurement study (or longitudinal design) is a focus on instructional characteristics affecting learning within the time period bounded by the time points of measurement – namely, the beginning and end of the selected grade of school.

Focus of a cross-sectional study – Most IEA (International Evaluation of Educational Assessment) studies have had a cross-sectional design, corresponding to fundamental interest in assessing and comparing cumulative achievement in a subject area as attained by students when they reach a given grade or age. The content domain has been defined as the essential core of the subject area as it accumulates and culminates at the target grade or year. We want to know in detail what students know and can do. The explanatory variables are personal (what kinds of students have what levels of attainment) and systemic (what school and teaching circumstances lead to higher achievements). The educational policy inferences are general and structural: for example, they deal with curricular arrangements, tracking and specialisation, retention and yield.

The effects of education are strongly cumulative and all of the learning that have been attained by a particular age or stage of schooling are a consequence of what has gone on before. Only if longitudinal study designs are used is it conceptually possible to isolate effects that occur within a specified, grade-year-limited time period and even this is a methodologically difficult task. As a consequence, explanations of cross-sectional data on achieved learning must rest on events that have occurred during the earlier as well as the contemporaneous process of schooling.

Focus of a longitudinal study – The fundamental interest is in the processes and connections of teaching and learning in the content area at the target grade level. The content domain is defined by what is being taught at the grade level. We want to know in detail about cognitive growth, about what students learn during the target year. The explanatory variables are the student input characteristics (prior attitudes and prior knowledge) and grade-specific measures of curriculum and pedagogy (content opportunities, teaching and classroom practices, teacher knowledge and student activities). The educational policy inferences are specific and formative: for example, they deal with the appropriateness of different content domains, with pacing and differentiation of content, with teaching methods and with student work.

Adapted from Wiley and Wolfe, 1992, pp. 297–300

Activity 6

Consider these tasks:

- *If you had to design a national assessment in a specific content area for your country, for a particular level of schooling, which survey design would you choose?*
- *Justify your decision in relation to the characteristics of the particular country, especially its needs, resources and the general context of education.*

Standards

The range of interpretations of the term 'standards'

It is often heard nowadays that 'standards are falling'. What are standards? Are they synonymous with national examination results, so that if more students do poorly on a test compared to a decade ago, we can say that the standard has fallen over time? They are also referred to as 'benchmarks of excellence' – could it then be the criteria used by examining bodies to allocate grades? Thus, only raw scores of 80 and over will be deemed worthy of an 'A'? Or, on the other hand, standards may not be related to final events like exams and grading procedures, but may actually refer to the teaching and learning processes in classrooms – are the expectations of

teachers and the methods they employ conducive to learning for the majority of students?

In most developing countries there has been expanded provision of education over the last few decades. Many will charge that in expanding provision, standards have fallen. If a country is attempting to provide schooling for all of its primary and secondary school population, then it is logical to assume that a wide variety of ability levels and background factors will be interacting with teaching and learning, and generally there may well tend to be a high failure rate. How then must we make sense of the charge here that standards have fallen from what they were a few decades ago? It seems safe to say that the sense in which standards have fallen is related to the inadequacy of the system in effectively teaching students with a wide continuum of aptitudes, abilities and background characteristics. There is a general feeling that schools are not as effective as they used to be and that this is reflected in a downward spiral of under-achievement.

The situation also seems to be clouded by value judgements about the way in which education is being transmitted. Many people have definite ideas about what is a sound education and what are important and valuable areas of study. Methodologies that involve student centred learning may seem time-consuming and lacking in rigour, given the demands of the examinations students have to sit. When improvements in curricula and innovative teaching methods do not seem to result in widespread higher achievement, there is a call for 'back to basics'.

This essentially means a concentration on literacy and numeracy, emphasis on drill and repetition and constant testing and monitoring of 'standards'.

The notion of education as a journey into a world of new and exciting experiences, involving the full participation of the student, retreats before a notion of education that is highly circumscribed by what ought to be taught, how it should be taught and 'accountability testing'. When one is trying to ensure accountability in the education system and sees that task as a strict maintenance of standards, the notion that not all standards can be adequately tested by the means we have at our disposal, or, even that 'standards' is an ambiguous concept, is inconvenient. Box 8 below explores at least three different understandings of the term 'standards'.

Box 8

'Learning' and 'Education' Standards

The World Conference on Education for All recognised that people's learning needs are very broad and varied: they 'comprise both essential *learning tools* (such as literacy, oral expression, numeracy and problem solving) and the basic *learning content* (such as knowledge, skills, values and attitudes) required by human beings to be able to survive, to develop their full capacities, to live and work in dignity, to participate fully in development, to improve the quality of their lives, to make informed decisions, and to continue learning' (Article I of the *World Declaration on Education for All*). In light of this, a 'learning standard' may be defined as a specific set of 'learning tools' and/or 'learning content' in the senses indicated above. In principle, it is understood that there exists a procedure for ascertaining whether a person has attained any given learning standard.

An 'education' standard may be defined narrowly as a learning standard that an educational programme aims to help learners attain or more broadly and colloquially as encompassing both the 'learning' standard aimed at and the whole complex of characteristics of the educational programme, for example class sizes, teachers' qualifications, textbooks required, conditions of the physical plant, and so on, associated with the standard. In this broader sense the term 'education standard' is virtually synonymous with 'educational quality'.

It is useful to draw a distinction between a prescribed learning or education standard on the one hand and a standard of performance or attained standard on the other. That they do not always coincide is reflected in the frequency of public complaints of 'falling standards', when the reference can be understood narrowly – depending on the context – to refer to level of learning achievement or performance of students in a particular educational programme or, more broadly both to this fall in learning achievement as well as a deterioration in the material conditions required to implement the educational programme under consideration. In this broader sense 'falling standards' is virtually synonymous with a 'decline' in the 'quality of education'.

Source: Adapted from UNESCO, 1993, p. 78

Activity 7

Consider these tasks based on the previous and following boxes:

- *Conduct a small survey of educators and members of the public to determine their notion of 'standards' in education. Do they equate standards with: achieved levels of proficiency as handed down by public examinations? Or is there more of an understanding of standards as a set of prescriptions detailed beforehand that students should attain? Do they subscribe to a broader notion of standards, as in the determinants of educational quality? Or is there some other understanding of 'standards'?*
- *Identify the sense in which the term 'standards' is used in Box 9 below.*

Box 9

Average Performance of Students in a Curriculum Area

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

This information is of limited value in itself because it does not tell us whether the average obtained can be regarded as 'satisfactory' or 'unsatisfactory'. It can be useful, however, if comparative data are available with which the obtained average score can be compared. Thus, the information could be useful as an indication of whether standards in the country were, in general, stable, rising, or falling, if comparable information were available from an earlier point in time. It could also be useful if similar information were available from other countries, as is the case in international studies of assessment.

Source: Adapted from Greaney and Kellaghan, 1996, p. 63

Summary

Putting this together, we come up with an understanding that all the senses of the term, 'standards', are related. When developing curricula, educators are mindful of the general prescribed minimum levels of competencies that world organisations and or national development plans lay down as desirable to upgrade the human resources of a country or region. In turn, these competencies guide the decisions of public examination and national assessment bodies in how they create test items, mark and grade scripts. And, certainly these 'minimum competencies' can be better transmitted in classrooms equipped with trained personnel and up-to-date resources. Hence, if persons only identify one 'sense' by which they understand 'standards', it is obvious that that implicitly invokes the other 'senses' with which 'standards' are associated. As an example, refer to Box 10 where there are regional prescriptions of standards, expected to be attained through the specific and general objectives of curricula and examinations. If there is a 'gap' between what has been prescribed and what is attained, then one can reasonably question the 'quality' of education being delivered.

Box 10

Defining Educational Standards in the Eastern Caribbean

Quality education is broadly defined in terms of educational standards expected of students at the end of secondary schooling. They should:

(a) Possess literacy and numeracy skills that allow them to:

(i) read with comprehension newspapers or magazines; health, disaster preparedness or agricultural bulletins; (ii) write legible letters in standard English to a friend, prospective employer or Government bureau; (iii) express views in an articulate and logical manner; (iv) handle important basic computations in everyday financial transactions;

(b) Know how to find information through the use of libraries, directories and encyclopedias, maps and charts, and modern electronic devices;

(c) Demonstrate positive habits and good interpersonal skills in the interaction with family, community, visitors and in group relationships generally;

(d) Demonstrate reasonable understanding and appreciation of scientific and technological processes as these pertain to nature, the environment and everyday life;

(e) Demonstrate a working knowledge of and functional capacity in at least one foreign language;

(f) Have an appreciation for a wide range of music, art, dance, dramatic expressions, and have some practical competence in expressing oneself in at least one of these areas;

(g) Possess functional knowledge and skills to facilitate entry into the world of work or to continue formal education;

(h) Possess functional knowledge and skills needed for civic participation –

  (i) as citizens of a democratic state,
  (ii) as participants in civic organisations,
  (iii) as members of the Caribbean Community;

(i) Hold to a value system that espouses a noble vision of Caribbean society and of Caribbean people and high principles pertaining to personal integrity, honesty, truthfulness and goodness.

Source: OECS, 1991, pp. 71–72

Activity 8

Consider this task and these questions:

- *Which of these standards are relatively 'easy' to test via traditional curricula and assessment practices?*
- *Research the topic 'Educational Objectives'. Identify the standards above, which lie in the 'affective' domain.*
- *It is possible that some affective objectives can be tested via 'authentic assessment'. Research this topic and evaluate to what extent authentic assessment is used in schools or public and national assessments.*
- *How valid is the proposition then that educators or the public in general can pronounce on 'falling standards?'*

Prescribed standards in specific curriculum areas

It is only in the specific curriculum area of literacy that there have been extensive international efforts to come up with a common standard of a literate person and translation of this into nationally prescribed standards at every stage and level of schooling.

> *A person is functionally literate who can engage in all those activities in which literacy is required for effective functioning of his [her] group and community and also for enabling him [her] to continue to use reading, writing and calculation for his [her] own and the community's development.* (UNESCO, 1993, p. 79).

Each country can either devise standardised tests to determine to what extent they are achieving this prescription and/or, they can participate in international literacy surveys so that they have a good idea of where they stand compared to other countries against the standard.

Millward, below, objects to the rigorous institution of standardised testing of reading in schools in the United Kingdom. The purpose of such tests is to ensure accountability – that the education system is delivering the minimum standards prescribed. However, the process of reading is multi-dimensional and testing can only touch few aspects. Emphasis on devising, monitoring and evaluating standards can unnecessarily circumscribe education.

Box 11

Standards and Basics

Talk of standards implies carefully defined benchmarks which can be used to describe actions and responses as being correct or incorrect. Standards are decentred, they demand consistency over time and they are applied indiscriminately across a population. Progress against a standard is marked in terms of moving closer to the norm or more nearly meeting the criteria. There are elements of the reading process which can be standardised. It is possible, for example, to identify carefully defined benchmarks (dictionary definitions of words or grapho-phonic correspondence) and to record correct and incorrect responses (right answers to comprehension exercises or correct pronunciation). It is possible to treat texts as constant over time ('What is Shakespeare saying here?') and not difficult to set out a hierarchy of features (moving from, say a recognition of initial letters, to vowels, to blends and diagraphs to strings of letters and so on) which describes and prescribes progress.

It is possible to describe elements of the reading process in this way, but it is immediately apparent that these benchmarks are not like the standard for measuring railway tracks.

It might be argued that we are only finding our way into describing standard features of the reading process and that, with time and with good teaching, we could expect to refine these benchmarks and make the standards more reliable. The important thing, so it might be said, is that we have identified significant aspects of reading (mostly concerned with decoding skills and the ability to say what a list of words or phrases mean) and that proper teaching of these skills will not only make children into proficient readers but also make the assessment procedures more reliable. After all, it is reassuring to know what to teach and reassuring to know that you have the procedures for deciding whether people have learned their lessons well. It is not hard to see how standardised testing can lead to children being taught the 'basics', and it is more than just a happy world which manages to ensure that the basics in reading are those elements of the reading process which are most easily tested.

However, this approach leads to a sterile account of reading with emphasis upon decoding, word building and 'given' meanings, and it is a description of the reading process which is severely limited. It is also a distortion. That the account represents only parts of the reading process for so long as it is made clear that only aspects of reading are being tested and for so long as there is 'no pretence that the results will tell a teacher or a parent all there is to know about the child as a reader'. However, the data recovered from standardised tests are usually treated as a measure of reading (consider only the implication of giving a child a reading age of, say, 7.9 years) and the standardised elements of the reading process are thereby presented as the essence of reading. That does matter. As readers' performances against these *standards* are seen to fall, it is easy to see how remedial action might be directed towards the *standards*. The 'basics' for success is realising standard performances. It is also easy to see how these *standards* can encourage teachers to take a narrow view of the reading process, a view which is reflected in their teaching and which colours the children's view of what counts as success in reading. This may bring disappointing results, but it helps us to see why the outcomes of reading tests are only rarely unexpected and it points to one of the reasons why the testers are able to claim high levels of content validity for their tests. We are in danger of being knotted up in a closed circle and condemned to a spiral defined by *standards* and 'basics'. Within such a spiral there can be no room for the reading process and it can lead to a curriculum which, 'through the assessment

framework emphasises the ability to conform to conventional correctness, and to produce prescribed responses'. The focus is 'almost entirely on a narrow conception of word identification and on children spotting 'correct' meaning in texts. This may be a narrow and dispiriting view of reading, but it lends itself readily to testing.

Source: Adapted from Millward, 1994, pp.88–89

Activity 9

Consider the following issues:

- *Should there be one literacy standard for all students at a specific grade level, or several to accommodate individual differences?*
- *Are there other, possibly more meaningful ways, of setting standards? Investigate different approaches to setting standards (Tuijnman et al., 1994).*
- *Identify the ways in which the setting, monitoring and evaluation of standards are a 'values-laden' exercise.*

International evaluation of educational assessment

When a country embarks on national assessments of its education system in the interests of greater accountability, it often has to become involved as well in international comparisons to obtain a more complete picture of what the system is achieving. What exactly will be compared? A score of some kind measured against a standard. National assessments need to be compared to past internal assessments to monitor 'standards' in the country over time, as well as compared to the performance of other countries to determine whether an acceptable level of performance is being achieved at home.

The IEA (International Evaluation of Educational Assessment) has undertaken many cross-national surveys where tests are made as equivalent as possible so that comparisons can be meaningful. The IEA also extends assistance and technical expertise in other ways of assessing the system – the emphasis is not only on testing. We are going to look in some detail at

the experience of the Dominican Republic (Box 12) in carrying out assessments of its education system with the help of the IEA. You will notice that both qualitative as well as quantitative data are collected so that the notion of 'standards' that is being invoked here has to do with data derived from tests as well as from qualitative factors having to do with the conditions and context of learning.

Box 12

Dominican Republic: The Study on Teaching and Learning of Mathematics

Most policy-makers do not know the educational achievement scores in mathematics and language attained by students spending a few years in the classroom. Therefore they do not know if the quality is better or worse over time. Unfortunately, many Latin American countries still lack national evaluation systems to determine curriculum areas that need improvement – systems that can generate innovative and viable programmes to improve instruction and learning in the various curriculum subjects. In fact, it is often left to the teacher alone to determine whether a student should be promoted to the next primary grade. Under this system, how can one be sure whether standards in one school are the same as standards in another?

In 1980, a group of professors initiated the TLMDR (the teaching and learning of mathematics in the Dominican Republic) project. Dominican professors decided to pool their knowledge in different areas to focus on a neglected but important aspect of education – the evaluation of classroom teaching and learning. One reason for interest in classroom mathematics teaching/learning was the perception that educators did not know how well scientific and technological knowledge appropriate to Latin America was being produced in the classroom. The lack of sufficient Dominican Republic college students interested and competent in the sciences could also be traced to poor teaching of basic sciences and mathematics in primary and secondary schools. In fact, educators had very little reliable information concerning the quantity and quality of mathematics learned by students in the Dominican Republic. Moreover, very little was known about the way teachers taught daily mathematics classes.

It was against this background that the IEA International Mathematics Study provided the opportunity to share access to expertise, experience and technological know-how accumulated by the IEA since the late 1960s. National expectations for the study were to receive empirical data which could be used to design new strategies to improve the teaching and

learning of mathematics in the Dominican Republic. The strategies would be based on information obtained from national samples of both teachers and students interacting with each other in the instructional framework. Comparisons with other education systems, both similar to and different from the Dominican Republic, could show how well students were doing on the basis of 'the international yardstick'. The earlier study (1980s) paved the way for participation in the newer international mathematics and science study now being conducted in over fifty countries by the IEA.

Like all international comparative studies, TLMDR was conceived as a broad-based, comparative investigation of the mathematics curriculum as prescribed, taught and learned. For the purposes of the study, the mathematics curriculum was seen to consist of three dimensions: intended (official documents), implemented (classroom teaching) and attained (student achievement gain on a pre- and post- test design). In addition to the curriculum analysis, the design allowed the study of several non-school factors, such as family influences, since it is always more appropriate to study an education system within its social context.

TLMDR mathematics tests were administered at the beginning and end of the school year to a random sample of 5,342 students in the target populations (13 and 17 year olds). Achievement tests consisted of a 40-item core and 4 rotated forms of 35 items each, from which students answered the core items and one rotated set of questions. In addition to the mathematics test, a student questionnaire was developed to gather information about parental occupation and education, students' nutritional practices, the occupational expectations of students and attitudes toward school and learning. Similarly, a questionnaire was developed for teachers, requesting information about academic training, teacher workload, instructional materials and practices, opinions about the mathematics process and attitudes towards mathematics. A classroom process questionnaire was developed to gather information about methods and procedures used to teach algebra, geometry, measurement, etc. Finally, a questionnaire was developed which gathered information about the opportunities students had had during their current or previous school year to learn the mathematics needed to answer the test items.

Some findings from the study: all school types in the Dominican Republic were below the international norms, with the private school students performing best among the school types in the country. In fact, the pre-test average score in the private school group was higher than the post-test score in all the other school forms.

Source: Adapted from Luna, 1992, pp. 448–51

Summary

From the above account the following is evident:

- Tests constructed for public examinations cannot be used for the purposes of national assessments although both are created with a similar attention to rigour.
- A national assessment extends far beyond the computation of raw scores and comparison with an international standard.
- National assessments are conceived as multi-faceted exercises where the opportunity is taken to gather as much data as possible on classroom as well as non-classroom factors.
- The notion of standards evident here is the setting of minimum competencies for particular age and grade levels and then comparison across international contexts.
- Low performance measured against the standard should signal that improvements are necessary in the system.
- Given the subjectivity, ambiguity and 'narrow' interpretation associated with the selection, monitoring and evaluation of standards, educators and planners conducting national assessments should be aware that accountability issues cannot be based on quantitative data alone.
- Standards can be regarded as a general indicator of system performance – the host of other contextual data collected indicates awareness that analysis of the system depends on a great variety of factors and standards are just a tool to indicate whether something is wrong.
- National assessments can reveal equity issues entrenched in the system, for example, performance of rural vs. urban schools, private vs. state-owned, denominational vs. non-denominational, or curriculum options that seem aligned to gender, ethnic or socio-economic factors.
- National assessments can also reveal if state funds are being used efficiently. Rapid turnover of teaching staff, high teacher pupil ratios, increasing rates of repetition, attrition and drop-outs as well as a large percentage of students being unable to obtain minimum credentials at the end of secondary schooling signal that investments in education are not realising targeted goals.
- The 'quality' of education is also revealed through the various data collection devices of national assessments. Curriculum analysis becomes a

must when large numbers of students are failing to attain minimum standards. Teacher training and classroom strategies also come under scrutiny as well as the calibre of texts and other resources. Teacher opinions are invaluable in pinpointing problems in the system.

Wrap-up and extended reflection

Launching a revolution in standards and assessments

The point is that standards will be meaningless if students continue to be tested without regard to them. Unless current tests change, the standards will wither and die. Teachers know that they will be judged by test results, and they will continue to teach to the tests by which they are judged. Tests must test knowledge and skills that are meaningful to students. The tests should be based on what children learn in class, and what they learn in class should be worth knowing and doing and using.

Standards are the starting point of education reform. You cannot design an assessment unless you have agreement about what children should learn. In the absence of national standards, we have evolved a haphazard, accidental, disconnected national curriculum based on mass-market textbooks and standardised, multiple-choice tests.

Education reform must begin with broad agreement on what children should learn. Learning, after all, is the heart and soul of education. When there is no agreement regarding what students should learn, then each part of the education system pursues different, sometimes contradictory, goals. As a result, our education system is riddled with inequity, incoherence, and inefficiency.

In the summer of 1991 the department sponsored a meeting with education ministers from 15 Asian Pacific countries. The subject of the meeting was standards for the 21st century. The United States was one of the few participants that had not established national standards. When we asked the various ministers why their society had set national standards, they all gave the same dual answer: to promote equality of opportunity and to promote high achievement. As they explained it, without explicit standards some schools would set high standards and others would not (and without assessments, no one would know the difference). Moreover, explicit standards enabled students to judge their performance against clear goals, thus raising everyone's expectations.

Ravitch, 1993, p. 772

Suggested reading

Anderson, S, Ball, S and Murphy, R (1975). *Encyclopedia of Educational Evaluation*. NY: Jossey-Bass.

Greaney, V and Kellaghan, T (1996). *Monitoring the Learning Outcomes of Education Systems*. Washington, DC: The World Bank.

Lockheed, M (1996). 'International Context for Assessments' in P. Murphy *et al.* (eds.) *National Assessments – Testing the System*. Washington, DC: World Bank.

Luna, E (1992). 'Dominican Republic: The study on Teaching and Learning of Mathematics', *Prospects*, XXII (4), pp. 448–54.

Millward, P (1994). 'Making Sense of Standards in Reading', *Evaluation and Research in Education*, 8 (1&2), pp. 85–96.

Murphy, P, Greaney, V, Lockheed, M and Rojas, C (eds.) (1996). *National Assessments – Testing the System*. Washington, DC: World Bank.

Nwama, O C (1996). 'What are national assessments and why do them?' in P. Murphy *et al.*, (eds.) *National Assessments – Testing the System*. Washington, DC: World Bank.

OECS (Organisation of Eastern Caribbean States) (1991). *OECS Education Reform Strategy*. St. Lucia: CIDA.

Ravitch, D (1993). 'Launching a Revolution in Standards and Assessments', *Phi Delta Kappan*, 74 (10), pp. 767–72.

Tuijnman, A and Postlethwaite, T (eds.) (1994). *Monitoring the Standards of Education: Papers in Honour of John P. Keeves*. Oxford: Pergamon.

UNESCO (1993). *World Education Report, 1993*. Paris: UNESCO Publishing.

Wiley, D and Wolfe, R (1992). 'Major Survey Design Issues for the IEA Third International Mathematics and Science Study', *Prospects*, XXII (3), pp. 297–304.