

The Place of National Testing in Educational Development: the Faces of “Strong” and “Weak” Assessment

H. Joseph Murphy, Dalhousie University,

Halifax, Nova Scotia, Canada

and

Andrew S. Hughes, University of New Brunswick,

Fredericton, New Brunswick, Canada

(A Background Paper Commissioned by the International Bank for Reconstruction and Development (the World Bank) for the Caribbean Education Seminar, Trinidad and Tobago, April 1999)

EXECUTIVE SUMMARY

There is considerable agreement among educators that national assessment and testing is capable of meeting both summative and formative goals. The summative ones include certification, selection and placement; and the formative ones include improving teaching and learning, curriculum auditing, monitoring of national achievement, and improving curriculum development. Illustrations from Caribbean countries as well as from countries outside the region are employed to show that the summative purposes are generally realized but evaluators and policy makers have not done a particularly good job in providing teachers and schools with the kinds of information from these national testing programs that are needed to realize the formative purposes. To increase the likelihood that both the formative and summative purposes of national testing programmes are realized, we make a case for considering selective sampling of participants in the testing, for exploring the alternatives to paper-and-pencil testing formats, for combining school-based teacher assessment and that derived from national examinations, and for improving the match between the form and use of the results of national tests. “League Tables” may be quite appropriate for parents who are seeking the best place to enroll their children, but they do precious little to help teachers improve the educational offerings in their schools; often they are left simply with the unhelpful exhortation to “try harder.”

We believe that it is possible to make national assessment and testing worth the considerable time and energy devoted to this activity. Several scenarios, derived from a consideration of work done in the Caribbean and in other countries, are described that go beyond assessment for certification, selection and placement to using results to improve teaching and learning, to audit curricula, and to monitor national performance. It is possible, through collaborative efforts among Caribbean nations, to build “strong” national testing programs that do attend to both formative and summative goals.

INTRODUCTION

We suspect that most of us here today know all about examinations. No doubt we have all had to sit more than a few. My colleague, Dr. Hughes and I can remember as children being subjected to batteries of intelligence tests that were supposed to identify our “learning potential.” If they ever did discover what that potential was we don’t imagine that the test ever did very much to help us either live up to our potential, or indeed, extend it. Most people don’t have very positive perceptions of examinations and tests. And yet, try to imagine a world without them. Let us tell you that we found it very assuring to know that the pilot of our flight on the way to Trinidad had passed his flying test — with flying colours, we hope.

According to the historian, Philip DuBois, a primitive form of proficiency testing existed in China as early as 2200 B.C. Not very much is known about this Chinese testing program other than that it involved some form of examination of public officials every third year. Much more is known about civil service examinations a thousand years later and which lasted until the early part of the twentieth century. Chinese candidates for civil service positions were evaluated for proficiency in music, archery, horsemanship, writing and arithmetic. Additionally they were examined with respect to skill in the rites and ceremonies of public and social life, civil law, military affairs, agriculture, revenue, and geography. There were three examinations:

- (i) Candidates composed a poem and wrote on an assigned theme while confined for a day and a night in a small hut used for test administration. The number who “passed” was usually between one and seven percent.
- (ii) Every third year those who had passed the first test were given a second and more rigorous one that required three days and nights of testing. Penmanship didn’t count at this level as official scribes copied the examinations before they were submitted to the examiners so as to reduce bias in marking.
- (iii) The one to three percent who passed test two were invited to the capital, Peking, where they completed a third test, with three percent becoming eligible for public service.

So, thousands of years ago, those responsible for national testing in China had worries just like those of test developers today. Do the tests have any validity? And how can we get rid of bias? Our focus today is going to be on questions of national testing and assessment. We want to paint for you a picture of assessment practices that can be found here in the Caribbean and elsewhere. We want to tell you that most national testing and assessment programs today are much like the intelligence tests that we experienced as children. They might well tell the assessors something but they do little to enhance and nurture better teaching and learning in our schools.

PURPOSES

In this section we would like to share with you some of the claims that are made concerning the virtues of national assessment and testing. We say “claims” because we are often not quite able to live up to the standards that we set for ourselves. In fact, there can be little doubt that we have been able to see much more potential in the realm of national assessment and testing than any of us have been able to deliver. Consider the following:

In Anguilla, it is claimed that the Test of Standards are to

- measure individual pupil achievement (strengths and weaknesses)
- identify students with special needs
- inform decisions on teaching
- monitor curriculum implementation
- determine the placement of students in secondary school
- compare performance of schools against national norms.

In Belize, the Belize Junior Achievement Test (BJAT) offers feedback to the education system concerning the development of numeracy and literacy skills at the end of the middle division of primary education. The Belize National Selection Exam (BNSE) serves largely as a selection and certification exam at the end of primary education. Beyond selection, results are used as the basis of a “performance analysis” for the education system; monitoring for the purposes of school licensing; and assessing and monitoring effective implementation of the approved curriculum. Furthermore, the information collected from the national exams is used to inform decisions concerning school system development and management through the deployment of available resources.

Since 1993 Bahamian High School students have completed the Bahamian General Certificate of Secondary Education, a locally developed examination designed to recognize positive achievement, taking into account varying and different abilities, rather than relative levels of achievement

Dominica is in the process of establishing national assessment procedures to commence in the year 2000. The general purpose is seen to be the improvement of teaching and learning. More specifically, there will be a focus on measurement of individual student achievement; the monitoring of curriculum implementation; and the setting of realistic standards.

Granadian students at the Grade Six level complete an 11+ examination that serves to select those entering secondary schools. Additionally, it is claimed that the results are used to inform schools about areas for improvement. Grade Four pupils will also be tested in 1999 with the intent of providing diagnostic feedback to schools.

Students in St. Lucia complete criterion-referenced examinations that claim to be either summative (Common Entrance Examination - Grade 6; Common Middle School Examination - Grades 8+9; and the Standard Six School Leaving Examination - Grade 8) or formative and diagnostic (Minimum Standards Examinations at Grades 2 and 4).

Compare these statements with some from outside the region. In Canada, we claim that there is a strong emphasis on the formative use of test results. Such uses can involve determining the strengths and weaknesses of the instructional program; adjusting

instruction to accommodate individual needs and monitoring student learning over time; and ensuring that each student receives the best education possible. The data is also used to identify areas of need for preservice and inservice education for teachers; to provide direction for needed research; to assist in the development of testing instruments for teachers; the monitoring of system performance by comparing schools and jurisdictions with available provincial and national information. System-wide examinations at the end of secondary school are administered in core subjects in four Canadian provinces. Results are combined with school-based marks to determine final high school grades that may be used subsequently for determining admission to higher education.

In France there is a system of “national mass diagnostic testing” at ages 8, 11 and 15/16 to: provide teachers with a tool to gauge students’ progress; assist teachers in selecting the teaching activities most suitable for the students’ needs; and to assist teachers in their planning of the teaching of the curriculum. In France, assessment is seen as a fundamental element at all stages. At the end of every year there is testing of selective samples to provide data for national and international surveys as well as for national monitoring and international comparisons. At the secondary level, assessment results are used to channel students into different types of courses and satisfactory completion of the high school examination (the baccalaureate) is the key to entry into higher education.

In England, compulsory baseline assessment at the time of school entry became compulsory in 1998. It has two main aims: to provide information to help teachers plan effectively to meet children’s individual learning needs; and to measure children’s attainment, using one or more numerical outcomes which can be used in later “value-added” analyses of children’s progress. It is claimed that the overriding purpose of National Curriculum Assessment is to contribute to raising educational achievement. At the secondary level, the GCSE is a summative examination. It is the main examination taken by nearly all students at the end of their compulsory school years. It is considered a useful predictor of potential achievement and, therefore, valuable in guiding decisions on post-16 education and training. Students seeking admission to higher education generally pursue the General Certificate of Education at the advanced level (A-levels). At the secondary levels, the examinations are generally purported: to raise standard across the whole ability range; to support improvements in the curriculum and in the ways in which it is taught; to provide clear aims for teachers and students, to the benefit of both and of higher education and employers; to record proven achievement; to promote the measurement of achievement on what candidates know, understand and can do.

In Singapore, the Primary School Leaving Examination and the Singapore Cambridge GCE examinations have strong summative dimensions being used to control access to secondary and higher education respectively. At the same time, there is an expectation of continuous formative assessment at the school level. The overt and explicit role of examinations in Singapore is similar to that in Bhutan where that national examinations serve primarily a selection function to determine access to the scarce number of places at the upper levels of secondary education and to provide a nationally accepted credential for employment. Of late there is a going recognition that the national examination results may serve other purposes such as curriculum auditing, improving teaching and learning, and monitoring performance at the school, regional and national levels.

What these illustrations show, and we believe that they are typical of most countries, is a range of generally shared interests. Most countries have an interest in the

question of selection. The task is that of properly identifying those students best suited to pursue particular types of higher education and/or employment, and/or of selecting the candidates most suitable for continuing education in systems where places are not available for all. The function here is strongly summative. At the same time, in most countries, there are claims of strong formative interests: improving teaching and learning, curriculum auditing, monitoring of national achievement, generating information for improved curriculum development as well as suitable professional development for teachers. But the fact of the matter is that large scale national assessment programs are not resulting in the sort of improvements claimed above. Clearly we believe that national assessments have the potential to go beyond the selection function, otherwise we would not make the claims, but there is little indication anywhere that we have been able to deliver on the promise. In spite of the widespread faith placed in national assessment programs in the role of educational reform in both developed and developing countries, we (meaning evaluators and policy-makers) have not done a good job in providing teachers and schools with the kinds of information that they need to improve teaching and learning, monitor performance, and enhance the quality of curricula in your countries and in ours. We have a common interest and a common problem that we can work on together. This is a point that we will return to later.

PARTICIPATION

In Anguilla, there is compulsory participation of all pupils in the national assessment program at grades three, five and six. In Belize, all primary students are required to sit the BJAT after completing standard three. All students completing primary are required to sit the BNSE. In Dominica there are plans for universal participation in national testing at the end of grades two, four and six. This sort of universal participation can also be seen elsewhere in the Caribbean as well as in Bhutan, Singapore, France and England. Such compulsory and universal participation in testing programs makes sense when the purpose is for selection, placement or certification, or, as in the cases of England and France, when data are collected to establish baselines and benchmarks for individual children. However, there are also instances when valuable and relevant data can be secured through less than universal participation. For example, there are instances of the use of representative samples as in Korea where the National Board of Education Evaluation makes assessments based on a stratified sample consisting of one percent of children in years 4 to 6, or as in the United States with the National Assessment of Educational Progress which is based on representative samples at the national, regional and state levels, and as in the cases of several Canadian provinces where samples of schools (Ontario) or pupils (Saskatchewan and B.C.) are employed. The point here is that valuable information can be generated for the improvement of educational performance based on samples of students but not when the principal interest is selection. Sampling is the appropriate means for efficiently obtaining valid and reliable data when the interest centres on:

- improving teaching and learning
- monitoring (auditing) curricula
- tracking national educational performance.

The benefits of sampling are obvious in large countries where the costs of universal participation in a range of evaluation activities can prove prohibitive. Sometimes they are

less obvious in small countries where the costs associated with sampling and its related administrative requirements may appear to outweigh the benefits. Here, the benefits of multi-matrix sampling may be employed. This might require universal participation but with each student completing only a sample of questions. In this way the focus is upon representative sampling questions rather than students (see attachment). While this procedure might initially appear complex and cumbersome, it is, in fact, straightforward and highly cost efficient demanding nothing in additional resources and expertise beyond that now available. (As any teacher will tell you, if you need a grade for every student then you have to mark all the papers; if you need a picture of how the class is doing generally, then a scan of a few papers will do the job.)

THE NATURE OF THE ASSESSMENTS

In Anguilla the focus of assessment in the national tests is on language, mathematics, social studies and science; the Bahamas employs tests in academic, technical and vocational subjects at the end of secondary school; in Belize, attention is given to mathematics, English and science or social studies; in Dominica, it is proposed that the focus should be on English (reading and writing) and mathematics; while Grenada assesses students in Natural Science, Social Science, mathematics and Language Arts. Attention to a similarly narrow range of subjects is evident in the National Assessment of Educational Progress in the United States, and in the various provincial testing programs in Canada. In Australia, attention in the primary school is upon “a national testing program for literacy and numeracy” with the intention being to assess the reading, writing, spelling and arithmetical abilities of children. At the higher secondary levels in most countries there is a tendency to conduct assessment in a wider range of subjects. In this regard, the assessed subjects of the Caribbean Examinations Council (CXC) are similar in scope to the English GCE’s, the French Baccalaureate and the German Abitur.

While assessment programs, particularly at the elementary levels, focus heavily on literacy and numeracy, examinations authorities in many jurisdictions are showing a willingness to explore the potential of “alternative assessment procedures” such as portfolio assessment, performance assessment, observation, consultation, self-assessment, peer assessment, anecdotal notes, notepads, reflective writing, checklists, markbooks, audio/video and photographs. However, it needs to be kept in mind that these are strategies recommended for on-going assessment by the teacher in the classroom. The emphasis in national examinations and testing remains upon essay or objective type tests, with the use of oral and practical methods when appropriate. Examination authorities appear keen to emphasize a balanced approach in the quest for assessment that is both valid and reliable. We have noticed that in the countries of the Caribbean there tends to be an emphasis upon the traditional, perhaps tried and true approaches. In Belize the pattern reflects multiple choice objective type tests and/or essays work. In Anguilla the emphasis is on paper and pencil tests though there is recognition for the need for some attention to oral and practical components.

Another important feature related to the nature of the assessment process concerns the relative weight given to school-based teacher assessment and the assessment derived from the formal examination. In most jurisdictions it is recognized that tests and examinations set and marked outside the school capture only a limited sampling of what

any student knows and is able to do. Continuous or on-going assessment by teachers is increasingly recognized as an important aspect of any balanced assessment of student performance. For example, in England teachers are required to use their knowledge of a student's work over time, including written, practical and oral work in the classroom, as well as homework, and the results of other school examinations and tests. In Singapore, continuous formative assessment forms an integral part of the teaching/learning process. Its purpose is to monitor students' progress in academic studies and extracurricular activities and it often involves such items as monthly quizzes. Data from this ongoing assessment is then combined with more formal testing to generate a summative assessment. In Victoria, Australia, the Certificate of Education is awarded on the basis of a combination of school-based and external examinations. In this case the weighting is two-thirds school based. While there appears to be a willingness to attach importance to the assessment made by teacher, there are, nevertheless, important related difficulties. Clearly, there is the problem of consistency of standards from one teacher to another and from one school to another. In high stakes assessment associated with selection there is subtle and sometimes not so subtle pressure applied to teachers from within and without their schools often resulting in what is euphemistically referred to as "the jacking up of marks." And there is also the problem of teachers being required to maintain complex and cumbersome records for which they have received little training. As far as we can tell, national assessment and school-based assessment have remained distinct enterprises for the countries for the Caribbean and these issues remain potential rather than actual.

THE FORM AND USE OF RESULTS

Belize represents an interesting case of the way results from national assessments are reported and used. For the BJAT (the tests at the end of standard three), individual student scores are not reported. Schools receive a report of the average performance of the school in relation to district and national means. Descriptive reports indicating areas of significant strengths and weaknesses are also provided. Each school receives its report only and school managers receive only those reports for the schools in their management area. While parents may ask to see reports they are generally unaware that they are available. Feedback from the BJAT is intended for use as a guide to teachers and school systems regarding the mastery of literacy and numeracy skills at the completion of the middle level of primary school. "BJAT results do not affect individual students in any significant fashion." On the other hand, the BNSE scores hold major implications for students. The BNSE is a high stakes examination as scores form part of the admissions requirements for secondary school. Poor performance on this exam may determine whether students are able to further pursue their education. While some students who fail may repeat the final year and re-sit the BNSE, most of the older students it marks the end of their formal schooling. Schools also receive a report based on the overall BNSE results for the purpose of providing feedback on the performance of the system. As with the BJAT schools receive feedback on their performance compared to district and national performance. In schools that continually show poor performance, funding is not affected but parents, teachers and students often suffer from low morale.

The two general procedures for shaping, reporting and using results, as employed in Belize, are typical of practices in many countries. We can frame the issues in terms of a set of questions:

- Are test results reported for individual students or as aggregated data sets for schools and districts? Or Both?
- What are the expectations concerning how results will be used?
- What are the implications for individual students and/or schools?

One of the most vivid uses of national testing program data comes from Great Britain. Results are reported to individual students and hold implications for admission decisions by the more desirable schools that are seeking to select “the brightest and the best.” On the other hand, the performance of individual schools receives scrutiny in the national press where “league tables” reminiscent of soccer or cricket league tables are routinely published, with some schools at the top of the tables and others at the bottom, perhaps even candidates for relegation. The intent is to provide a tool for parents making choices about the education of their children at the secondary level. The tables provide background and information on each school and show the achievement of students in GCSE exams at the secondary level with results at the primary school level focusing on English, mathematics and science at age eleven. Beyond this, the results of national curriculum assessment may also be used as a resource to help schools raise standards and assist their students to reach their full potential. Since 1997 schools have been required to set and publish annual targets for student performance on the national exam. It is intended that these benchmarks will be used by schools to set targets for improvement. Schools with already high levels of achievement are expected to set target for further improvement. All of this is couched in the spirit of what is being described as a “value-added” approach to education.

Other countries such as Australia have followed the British example. In the state of Victoria since 1996 there have been league tables incorporating “some attempt to measure value-added” to “celebrate achievement and provide accountability to parents.” In Singapore only the results of the top 50 schools are published. At the other end of the spectrum are systems that place little emphasis upon individual student results. In the province of Ontario in Canada, the Ontario Provincial Learning Assessment Program (OPLAP) is a curriculum-based approach that was developed following a study of existing large scale assessment models. The OPLAP which evolved from Ontario’s participation in the Second International Mathematics Study (SIMS) includes the collection of data on the intended, implemented and attained curriculum. The results are generated from a representative sample of one hundred schools per language group (English and French). The main uses of the program are for public accountability and the improvement of school programs. Obviously, results are not available for every student in Ontario so they do not play a role in selection and certification.

Whether a focus on individual student results or aggregated school and system results is appropriate depends largely on the intended use. Clearly when there is a need for selection, certification and placement, individual results are essential. But when the principal interest is in public accountability, curriculum auditing, assessing teaching and learning needs, and tracking national educational performance, then aggregated information based on representative samples of students is both effective and efficient. In spite of the rhetoric in many countries about the wide range of purposes to which testing data are applied, we know of little evidence that demonstrates that information from large scale national assessment activities are being used in a productive and meaningful way beyond the selection, certification and placement of students. In a recent

web-based discussion on major issues in assessment through Division D of the American Educational Research Association, one participant said “there is a lot on how they (large scale assessments) SHOULD be used, or how the CAN be used,” but, “does anyone know of any good evidence that large-scale assessment systems are actually resulting in improvements in education . . . in other words, how have people taken the assessment information and made positive changes in educational programs.”

So, what we want to emphasize here is that while national assessment is generally thought of as possessing considerable potential in positively influencing the development of education in all countries, there is little evidence that it has in fact done so, and there is some evidence that the sometimes considerable investments in national assessment have little impact upon the quality of education provided in schools. Indeed, the guidelines from the World Bank for the preparation of this paper say that “ weak monitoring and assessment systems remain major obstacles for improving learning outcomes.” And what we want to tell you is that weak monitoring and assessment systems are those that emphasize the assessment of individual students for selection, certification and placement at the expense of meaningful data that can be used to:

- specifically and directly improve teaching and learning;
- track national, district and school performance over time;
- monitor and support the implementation of approved curricula.

We would be prepared to acknowledge that the systems with which we are familiar, in our home country, and in many others, are by this definition “weak systems.” We would leave it to you to judge whether your country falls into the same category. But we want to stress is that the systems are weak, not because of the absence of good intention but because of the failure of assessment specialists and policy-makers to deliver on the potential and promises of assessment.

We have visited schools where the head teachers and staff have been provided with data in the form of standard scores and raw scores, percentile ranks local and national, grade-equivalent scores, benchmarks and key-stage levels of attainment. Or, at the other extreme, reams of computer print-out showing profiles and performance tables detailing the performance of every child on every question in every subject. The implication is that somehow teachers and heads will take this information and translate into an meaningful program of school improvement. It does not happen! It will not happen! It cannot happen! What we do as assessment specialists and policy-makers is download onto schools and teachers a responsibility that they cannot possibly discharge. What we need to do as policy-makers and assessment specialists is be clear about our purposes, generate information that informs the purposes, provide it to schools in a form that is meaningful and usable, and provide schools with the assistance that they will need to use the results in productive and purposeful ways.

MAKING NATIONAL ASSESSMENT WORTH THE EFFORT

In many countries, the agencies responsible for national assessment and testing are seen by educational professionals in schools as bureaucracies that suck up resources that could be better used elsewhere. How could they be viewed otherwise? They intrude upon the daily routine of teachers and students, providing little in the way of tangible benefits. The challenge confronting us now is to specify specific steps that can be taken

so that national assessment can fulfill its promise, contributing to the enhancement and development of education in all of our countries. So let us describe a few scenarios that go beyond assessment for certification, selection and placement.

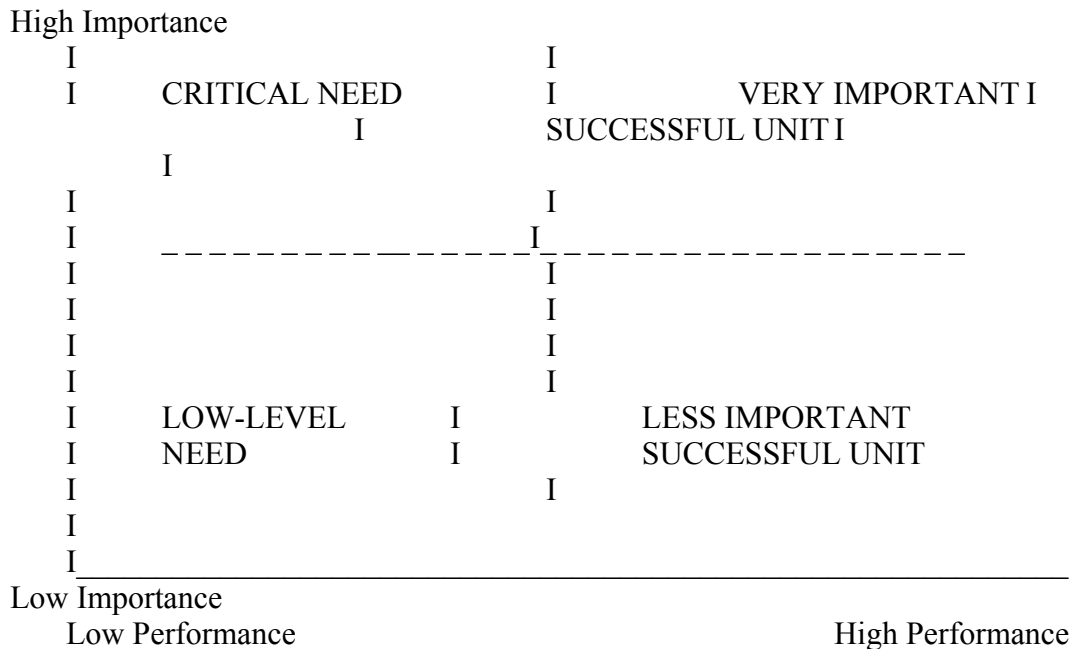
Using Results to Improve Teaching and Learning

Generally speaking the information provided to schools from national assessments comes in the form of a statistical profile of some type, usually a single mark for each student in each individual subject. Beyond calculating the average marks and the proportion of students passing, schools do not make much use of the data. Indeed, little use of results in this form can be made though teachers and heads often struggle with the information in an attempt to improve teaching and learning. It is a fruitless task given the information they are typically provided. At the same time the results of the assessment programs could provide the data that would permit a comprehensive instructional needs assessment at whatever level deemed appropriate, whether national, regional or even the individual school.

One approach would be to use the information to generate a teaching/learning needs assessment. Such an assessment would allow Ministries of Education and even individual schools to identify the strengths and weakness of students in any examined subject and thereby help to establish priorities for curriculum development and inservice training.

FIGURE 1

Teaching/Learning Needs Assessment: Matrix Analysis Template



In this approach, each course would be structure into units, chapters, objectives (outcomes) and data generated concerning the relative importance and students' performance in each area. The data are then entered on the template which is divided into four cells. Plotting the data into the matrix reveals which areas (units, topics, outcomes) are in need of attention. The key quadrant is the one labeled "critical need," which identifies those units or outcomes in a course that are viewed as being highly important but where student performance is relatively low. Such units or outcomes are clearly demanding of greater consideration than, for example, those which are of relatively low importance and where student performance is relatively high. The results of such an analysis can clearly be of value to curriculum planners and designers, managers and administrators in directing the sort of assistance that should be provided to teachers and schools. Data presented in this form at the level of individual schools allow teachers and heads to give focus to their own endeavours to improve teaching and learning and can be important elements in a school improvement plan. Our experience with this shows that teachers respond favourably to the approach. It is highly transparent, requires no particular technical skill, and can facilitate high levels of consensus among educators at all levels concerning the priorities to be pursued. The knowledge and skills necessary to analyze, present and interpret data using this approach can be readily acquired, and could be the focus of a short in-service training program.

Using Results to Audit Curricula

A major concern for school authorities throughout the world is whether the officially prescribed curriculum is being to the fullest extent possible at the school level. After all, it is the official curriculum, as reflected in policy statements and curriculum guidelines that specifies the range and depth of opportunities to learn that are to be provided to students. Curriculum Auditing (English, 1988) is a procedure designed to help curriculum managers monitor the quality of curriculum provided in schools. An important element of this task is comparing the performance of students (the Learned Curriculum) with the Official Curriculum and the Taught Curriculum. An important first element in this process can involve a comparison of the intended outcomes of a program with those actually achieved. This initial process can identify areas of significant discrepancy and was important part of the recently completed Third International Math and Science Study. An examination of the Taught Curriculum which could be the focus of site visitations by Ministry Officials (as in the case OFSTED officials in England and Wales) can then reveal whether such discrepancies are the result of a failure to deliver the official curriculum in the classroom or a failure of the official curriculum to achieve its intended outcomes. The results of the auditing process provide information relevant to the work of curriculum officials, those responsible for inservice training and development, and of course, for the work of individual heads and teachers. Similarly, some specific knowledge and expertise is required of those engaged in the auditing process. It is knowledge that is focused, to the point, and non-technical. It can be acquired readily through short-term training.

Using Results to Monitor National Performance

As ever increasing resources are devoted to educating children in all nations, national leaders inevitably ask whether students are better educated now than they were formerly. At the present time, most nations have little information that will allow the question to be addressed in any sort of systematic and valid fashion. And yet, information

from national assessment programs is readily adaptable to the monitoring of national educational performance with little or no additional cost or effort. What is demanded is the establishment of benchmarks against which further development can be measured. The data from current assessments can be used for this purpose. Results from subsequent examinations can be used to monitor the direction and amount of growth on a year by year basis. From this monitoring procedure, national trends can be identified. Since the test data is collected at the school level, trends can be observed for individual schools and aggregated into district summaries. Monitoring student performance on a year by year basis does require comparability of the testing instruments. There are standard procedures available to ensure such comparability, generally making use of anchor tests or items. Some particular technical expertise is required here but is necessary only for measurement specialists and not for school heads and teachers.

SOME PARTICULAR ISSUES RELATED TO ASSESSMENT IN SMALL COUNTRIES

In our discussion above we have referred to generally wasteful and inefficient procedures associated with universal testing when data from samples would serve the same purposes. However, we know that there are particular issues here related to the situation in small countries. During the past year we conducted a review of the national examinations used at the Class 6 and Class 8 levels in Bhutan. Some international agencies have been urging Bhutan to adopt a standardized testing program that would permit international comparison of educational achievement. These agencies have been suggesting that such assessment should be based on the use of representative national samples. Here is a country of less than a million people with some 6,000 children in class 6 and places for 400 in class 12 (the last year of high school). At the present time, there exists a set of high stakes examinations at the Class 6, 8, 10 and 12 levels that determine access to the next level. Selection and placement is a central interest and will remain so. Our own argument would support the notion of generating data from samples when the purposes extend beyond selection and placement. However, the sample sizes required to secure representativeness would be so large, in proportion to the total population, that the sampling process itself would become cumbersome, wasteful and inefficient. In this situation it absolutely makes more sense to require universal participation. And, indeed, this is the situation in most small countries.

Whether based on samples or universal participation, there remain the requirements of clarity of purpose, the design and development of valid tests, the efficient administration of the testing process, the appropriate and timely analysis of results, communication of the results in a manner that is meaningful to both educational practitioners and the public, and the provision of the support required to make purposeful and productive use of the information that has been generated. This is no small undertaking. It is not something that can be simply tacked on to the duties of some already overworked Ministry official. It is something that requires a cadre of trained personnel. It is an expectation that may lie beyond the individual capacity of many small states. Collectively, however, it is something that lies within the grasp of alliances of small states working together. We have already seen instances of this with, for example, the Caribbean Examinations Council, the University of the West Indies and the

Organization of Eastern Caribbean States. Working together, small states can ensure that their own needs are front and centre, that their own personnel develop the knowledge and expertise required, that assessment can be part of the indigenization of their systems of education. In spite of the implied promises of Educational Testing Services in Princeton or the Cambridge Examinations Board, national systems of assessment are not simple commodities. They cannot be delivered by FedEx in a box. This is, of course, an argument made often before in Caribbean circles, and elsewhere.

BUILDING “STRONG” NATIONAL TESTING PROGRAMS

Systems of national assessment can be efficient without being “strong.” Systems can be used to efficiently and effectively select, certify and place students without having the strength to nurture and direct further development. Systems that do this and only this are “weak” systems. “Strong” systems can, of course, attend to issues of selection, placement and certification, but they are not limited to them. “Strong” systems go beyond, playing a leading role in helping to improve teaching and learning, in monitoring curriculum, and tracking national achievement. “Strong” systems participate in nurturing talent; “weak” systems see assessment as a matter of selecting and sorting our most valuable resources, our children. School children and teachers throughout the entire Caribbean region would greatly benefit by efforts to help move the national testing programs from “weak” to “strong” systems.