Cambridge Assessment

Design challenges for national assessment in this accountability era

A background paper commissioned by Cambridge Assessment

Sandra Johnson

October 2017

How to cite this publication:

Johnson, S. (2017). *Design challenges for national assessment in this accountability era: A background paper commissioned by Cambridge Assessment.* Cambridge, UK: Cambridge Assessment.

As a department of Cambridge University, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

http://www.cambridgeassessment.org.uk

researchprogrammes@cambridgeassessment.org.uk

Contents

Executive summary

Stimulated by the activity and findings of the international survey programmes that now dominate educational policy-making worldwide, over the past 15 years or so governments around the world have embarked on reviews of every aspect of their provision, with policy initiatives following in curriculum, instruction and assessment. In particular, national or regional curricula have been newly introduced where none existed previously, and national assessment programmes have been newly launched, or significantly remodelled, in countries as far apart as Australia, Brazil, the Czech Republic, France, Georgia, Germany, Rwanda, Saudi Arabia and Sweden, among many others. Brief overviews of national assessment activity in the major world regions are offered in this paper to contextualise the scope and scale of continually evolving programme development.

There are a variety of officially stated purposes for national assessment. Principal among these will be the provision of information about the achievements of students in one or more age groups, or at the beginning or end of particular stages of schooling, currently and over time (system monitoring). National achievement levels in some curriculum-relevant subject domain will be the focus of interest – typically reading and numeracy in the primary sector, and language, mathematics and occasionally science in the secondary sector – along with relative subgroup performances (gender, socioeconomic grouping, school type, etc.). Descriptions of learning environments in school and at home are also often sought, along with attitudinal information about learning from teachers and students. Such information might be gathered simply to contextualise the attainment findings, or to provide data with which to establish statistical associations between 'background variables' and attainment, in the hope of establishing cause-effect relationships – a vain hope where cross-sectional data is concerned. Sometimes this data contributes to the impact evaluation of a recently introduced educational policy. A more sporadic purpose for national assessment might be to promote or even to accelerate curriculum reform, especially where this calls for new approaches to teaching and assessment on the part of teachers (exemplification). The most contentious, and arguably the fastest growing, purpose for national assessment is that of school accountability.

It is generally acknowledged that selection and certification systems that are high-stakes for students, (e.g., school-leaving qualifications), offer important challenges to designers and managers in terms of dependability, fairness, transparency and stakeholder buy-in. What is less widely recognised is that national assessment programmes, which can be high-stakes for teachers and schools, pose even greater challenges in terms of their design and implementation, particularly given their even greater vulnerability to political control and commitment. Many design choices are in principle available to programme designers in attempts to ensure that programmes meet their intended purposes, although not all will be available in practice because of political diktat, logistic constraints, and resource shortages, among other factors. Design decisions concern curriculum focus, age groups/stages, survey cycles and timing, modes of assessment, scale of assessment, underpinning measurement model, attainment reporting style, and stakeholder dissemination strategy.

Purposes and design choices are outlined in the paper, as are growing trends. The kinds of insoluble design challenges that are faced when purposes are multiple, non-prioritised and possibly even conflicting, are also overviewed. Finally, on the basis of reported experience on the part of some of the many individuals and organisations that have been directly involved in implementing national assessment programmes around the world, the kinds of difficulties that can be met in practice are identified, some of which have proved fatal for the programmes concerned.

1. Introduction

1.1 The emergence of national assessment

The history of national assessment for system monitoring is closely intertwined with that of the international survey programmes that now dominate educational politics worldwide. After more than half a century of activity, we have reached a situation where international surveys are taking place on a regular basis, touching every part of the world. For school-level systems monitoring we currently have TIMSS (now an acronym for *Trends in Mathematics and Science Study*) and PIRLS (*Progress in International Reading Literacy Study*), both run by the *International Association for the Evaluation of Educational Achievement* (**IEA**), not forgetting PISA (the Organisation for Economic Co-operation and Development's [OECD] *Programme for International Student Assessment*).

With a first survey in 2000, PISA was launched with the principal intention of filling an indicator gap by providing OECD countries with information about system output, specifically student achievement at age 15, to put alongside information about input (such as human and financial resources) and process (system organisation). The new availability of comparative information about system outcomes at the end of obligatory schooling in many countries, provided by such an authoritative body, could have been predicted to have an important policy impact internationally, arguably an impact far greater than TIMSS and PIRLS have achieved. But the OECD itself might have been surprised by the immediacy of the reaction, and by the extent and importance of PISA-inspired, or at least PISA-supported, educational reform over the past decade and a half. A number of countries, in Europe and elsewhere, experienced what has become known as 'PISA shock' when the first PISA results were published (OECD 2001), despite the fact that in many cases concern about national performance was unwarranted (Baird et al. 2011, 2016).

In response to those early cross-national findings, and those that have followed in subsequent surveys, governments around the world have embarked on reviews of every aspect of their provision, with policy initiatives following in curriculum, instruction and assessment. National curricula have been reviewed, as in England, in the primary as well as the secondary sector, even though PISA claims not to be curriculum-based. Common regional or national curricula have been newly introduced where none existed previously, as in Germany and Switzerland. National assessment programmes, too, have been newly launched or significantly remodelled in countries as far apart as Australia, Brazil, the Czech Republic, France, Georgia, Germany, Rwanda, Saudi Arabia and Sweden, among many others (Eurydice 2009; Rey 2010; Sargent, Foot, Houghton & O'Donnell 2013; Table 9.1).

By their nature, international attainment surveys are composed of synchronous national attainment surveys, which means that national governments benefit from information about their own domestic education systems when they participate in international surveys, in addition to learning about how their system compares with others (Kellaghan, Bethell & Ross 2011). Through the option of enlarged student samples, national governments are able, often for the first time, to gather information about their systems at within-country regional level as well as national level – Canada, Germany, Switzerland, and the US (some states) are among several examples of countries that have exploited this option. Why, then, should any country feel the need to launch or to retain domestic survey programmes in addition to participating in the international surveys? There are a number of reasons (Greaney & Kellaghan 2008; Kellaghan, Bethell & Ross 2011).

Firstly, participation in international surveys is expensive, particularly should student over-sampling and associated data analyses be requested. Domestic surveys will be costly, too, depending on their nature, but will likely be less expensive than international participation. Secondly, while the international surveys are now all conducted on regular cycles, the cycles

are relatively long: 4 years in the case of TIMSS, 5 years in the case of PIRLS, and 9 years in the case of PISA (for 'major domain' coverage). While such intervals would seem reasonable in terms of detecting any attainment trends that might reasonably be expected to emerge, national politicians might find them inappropriate for meeting their shorter-term system monitoring and evaluation needs. Thirdly, the international surveys focus on particular knowledge and skill domains, at selected stages in schooling, and might not address some key subjects in national curricula, or might address them only partially; indeed, PISA, by design, does not address specific curriculum issues at all, 'transcending' national curricula by addressing generic cognitive skills rather than specific subject knowledge, providing any needed knowledge within its test questions. Fourthly, when national initiatives are launched, for example, to stimulate an interest in science among lower secondary school students in the hope of raising their science attainment, focused domestic surveys are an appropriate way of evaluating impact.

For all of these reasons, and more, national assessment comfortably takes its place in a world dominated by international surveys, if albeit often promoted and financially supported by donor agencies, and in consequence influenced by international survey designs.

1.2 Principal and complementary purposes

The purposes of national assessment are today many and varied, as the examples in Section 4 will attest. The principal, explicit and shared 'complex' purpose continues to be a) to establish an 'attainment baseline' in given knowledge-skill areas at particular stages in schooling, for the given student populations as a whole and usually also for subgroups within these (gender, deprivation, language of learning, for example,), and b) to monitor change in the initial attainment profiles over time. This multi-faceted purpose – 'system description' and 'system monitoring' – often co-exists with a range of complementary ambitions. These include the potentially invaluable aim, as far as the interpretation of attainment change is concerned, of establishing relevant learning environments within and outside the classroom, including resource availability, teachers' instructional styles, teachers' and students' subject attitudes, and so on. A more controversial purpose for national assessment, and the one that holds irresistible and growing interest among politicians, is that of school accountability.

Information for system description and system monitoring indicates where nations stand, alone or in comparison with others, in terms of the overall achievement of target student populations. Indicators can include test and subtest mean scores, or, more usually at the present time, proportions of students falling into ordered groups in terms of their 'levels of proficiency'. It also includes information about the relative achievements of specific identifiable subgroups within the target populations, in particular boys and girls, groups distinguished in terms of socioeconomic background on some measure, regional groups, ethnic groups, rural versus urban groups, groups in different types of school, and so on. Attainment reporting will typically be by subject domain (e.g., reading comprehension) and subdomain ('information retrieval', 'comprehension', 'analysis', etc.), but might also, if rarely, be at a more detailed level, for example,, exploring strengths and weaknesses within subdomains.

While generally gathered simply to contextualise attainment findings, possibly *suggesting* potential links between attainment and conditions of learning, questionnaire-based enquiries into student demographics and learning environments are sometimes used to furnish data in the usually vain hope of *confirming* cause-effect associations. There is a temptation in this type of 'big data' enterprise to draw inferences about cause and effect, even in the knowledge that cross-sectional surveys cannot by definition provide cause-effect *evidence*. Longitudinal surveys offer greater possibilities for providing valid evidence for cause-effect relationships, but these are more difficult to implement and have been quite rare in sample-based contexts, particularly where paper-based testing is the norm. The growing popularity

of annual cohort testing in consecutive grades, such as is the case in Brazil, Chile, South Africa, Qatar, and Hungary, among other countries (see Section 4), does in principle provide an appropriate strategy for cause-effect investigation, especially where school systems are well-equipped to accommodate online testing. Costs, here, can be inordinately high, however, the logistic problems surrounding student testing formidable, and the effects of repeated testing of individual students potentially harmful.

In principle, national assessment programmes are designed to reflect the impact of a given school curriculum, whether in terms of achievement in key school subjects, such as science, geography or health and wellbeing, or in terms of generic 'across the curriculum' knowledge and skills development, typically literacy and numeracy. They can also be used to accelerate curriculum reform, for example, by using assessment tasks to focus the attention of teachers and others on the knowledge and skills, and sometimes the attitudes, that a new curriculum has been designed to develop in students, and illustrating ways that such development might be validly assessed. England, Scotland, and Chile offer just three examples among many (see Section 4 for details).

Arguably, the most controversial purpose assigned to national assessment is that of school, as opposed to whole-system, accountability. Requests for school-level information to be provided from the international surveys have been made by some country participants. However, never having been designed to provide this kind of information, the technical challenges associated with meeting such requests with any validity have been confirmed to be insurmountable (see Mirazchiyski 2013 for an account in respect of the IEA programmes). In the meantime, several countries and states have at some point introduced school accountability as a major, if not principal, objective of their domestic assessment programmes, in some cases with penalties or rewards for schools depending on their 'performances'. Examples include state assessment systems in the US, and cohort-based national assessment programmes in Chile, England, Hungary, Scotland (a very recent development), and South Africa (Section 4 provides detail for these and other country examples).

If not already planned from the outset, successive surveys at particular school stages in particular subjects will be organised, in order to monitor attainment over time in the hope of detecting short-term change or longer-term trends: thus, national assessment for system description becomes national assessment for system monitoring. Where policy initiatives have been implemented that are designed in principle to raise attainment, then stakeholder expectations will be that any discerned change in attainment will be positive (policy evaluation).

1.3 Stakeholder groups and programme control

Of the various stakeholder groups with an interest in national assessment and its outcomes – national politicians, policy-makers, educators, parents, and employers – it is generally politicians and policy-makers who make the key decisions that shape the scope and scale of surveys, and of assessment programmes as a whole, taking into account national needs but also inevitably influenced by international trends. National politicians and policy-makers have the greatest power and influence, in deciding whether and how national assessment should take place, what its purposes should be, what general form it should adopt, what financial and other resource support it should be allocated, how and how widely its findings should be disseminated, what would be an appropriate policy response to findings, and when programmes should be modified – 'rebranded' perhaps – or discontinued. While this power group will probably take advice from other stakeholder groups, and from experts in the assessment field, their national assessment decisions will also be unavoidably influenced, for better or for worse, by international trends.

Regional and local politicians and policy-makers will have an interest in the findings from national surveys, and could in principle use national findings to implement and evaluate local policy initiatives. These individuals might also successfully argue for specific regional issues to be addressed within national surveys, for example, making a case for regional oversampling of students in surveys, or for the introduction of regionally-specific questions within student, teacher, or parent questionnaires. Where a school inspectorate exists, its staff will have a professional interest in survey findings, about student attainment, teaching practices, learning environments, and so on. These individuals, too, might have opportunities to input to the design of surveys, including the nature and content of tests and questionnaires, and they would have a professional obligation to act on relevant findings where a need for improvement in the education system might be indicated.

Managers of education systems and school principals might have little if any influence on the design, or even on the very existence, of a national assessment programme, but they will have a professional interest in programme findings, particularly where these relate to individual schools. Teachers and teacher trainers, too, will have a professional interest in survey results, particularly in cases where programmes are in part designed to have an explicit or covert role in reforming the curriculum or changing teaching practices. Educational researchers, and in particular school effectiveness researchers, should enjoy the availability of accessible national-scale student attainment data, along with questionnaire information about learning environments as a potentially rich research resource. Parents, employers, the general public, and the media will view survey findings with interest; particularly should league tables be published.

For their part, students might have little intrinsic interest in what large-scale surveys have to say about their school systems, and yet they are the building blocks of the enterprise – without them, and their serious participation in the testing, attainment surveys would have questionable validity and little policy value. For this reason, some countries take special steps to explain to students the importance of the enterprise, if not for them personally then for those coming behind in the system, in an effort to increase survey participation and motivation.

1.4 Design challenges and trends

National assessment programme design is a challenging activity, even without the irresistible and frequently changing political pressures associated with it. Designs can be technically simple or highly complex. Implementation costs can be moderate to extremely high, depending on the scale of testing, and the scope in terms of curriculum coverage. The burden on schools can be minimal or excessive. And outcomes, in terms of meeting expressed political expectations, can be entirely satisfactory or intensely disappointing. Designed well, within the particular national constraints of geography, transport and communications infrastructures, financial resources, technical expertise, logistical practicality, political stability, and other critical factors, national assessment can be a powerful tool for system evaluation. Designed poorly, national assessment programmes can waste time and resources, whilst providing little information of genuine value to any stakeholder group. Sadly, lessons learned from failed models are rarely widely disseminated, but are learned afresh by others, when time has passed, money has been spent, teachers have been overburdened and frustrated, and information gained about system effectiveness has proved inadequate with respect to expectations.

This paper is primarily concerned with issues of programme design and implementation. Section 2 considers the emergence of national assessment half a century ago, and its continued growth in scale, scope, and ambition to the present day, influenced in developing countries by the capacity building efforts of international aid organisations supporting comprehensive education reform packages. In Section 3, the challenging issue of programme design is explored. In addition, some of the common implementation problems

that have been met in practice, and which have threatened the fundamental value of programme findings, and even the continued existence of programmes themselves, are identified. Section 4 brings the history of national assessment activity up to date, by offering accounts of practice around the world, illustrated with selected country examples. Finally, Section 5 offers reflections on global experience, identifying increasingly common trends in programme expectations and design.

2. Drivers of growth in national assessment activity

2.1 The early and continuing influence of the IEA

The emergence of the first national assessment programmes, in the US in the late 1960s, followed by the UK in the 1970s, was arguably triggered by the first cross-border surveys carried out by the IEA. The IEA was founded over half a century ago, in 1959, by a group of eminent school effectiveness researchers, from the US, the UK, Sweden, and elsewhere, with the intention of carrying out cross-border attainment surveys with an explicit research focus. National surveys had repeatedly provided evidence that student-based factors, especially socioeconomic background, strongly outweighed school-based factors in attainment correlation studies, so that it was difficult on the basis of single-nation research to provide any evidence that schools 'made a difference'. It was hoped that looking across borders would provide richer information with which to explore the issue of school impact (Husén & Postlethwaite 1996), since this would introduce greater heterogeneity into educational provision:

> *The world could be conceived as a huge educational laboratory where different national practices lent themselves to comparison that could yield new insights into determinants of educational outcomes.* (Purves 1991: 34)

From its earliest beginnings, the IEA has focused its primary sector surveys on 9-10-year-olds and its secondary sector surveys on 13-14-year-olds with occasional assessments of 17-year-olds. In each age group, in each participating country, nationally representative samples of students are selected for testing, by first randomly selecting a sample of schools (controlling for characteristics like size, location, and socioeconomic composition, i.e., by 'stratifying' the population of schools before making selections), and then by taking one class of students, or occasionally more than one, from the relevant year group in each school. In addition to taking tests, students also answer 'background' questionnaires designed to gather information about their learning circumstances at home and at school, about their subject attitudes and interests, and about their classroom experiences, among other topics. Teachers complete questionnaires, too, covering issues such as demographics, teaching experience and teaching style, the quality of subject resource provision, and curriculum coverage.

The IEA's first venture was an international collaboration in mathematics education, with 12 developed countries participating in the early 1960s in the first IEA attainment surveys in this area of the curriculum, at ages 13 and 17 (Husén 1967; Wilson & Peaker 1971). The form of those first surveys was already comprehensive, and clearly in line with the founding philosophy and intentions of the Association. Students' performance was assessed for a number of different aspects of school mathematics, with reference to an identified commonality in national curricula, and a large volume of information was gathered about students' circumstances of learning.

The IEA had no political motivation for its early survey activity. Its particular research focus nevertheless inevitably meant that national attainment outcomes were being incidentally exposed in survey reports (Peaker 1975). This will have whetted the appetite of national policy-makers for further outcomes information for their own systems, along with continually updated information about their standing internationally. But the IEA was not in a position in those early days to guarantee further subject surveys, least of all on a regular basis, since it depended each time on sufficient numbers of interested governments agreeing to contribute implementation funds. In response, some countries took steps to develop assessment systems of their own, to provide regular national outcomes information.

The US was the first country to introduce a formal national assessment programme when it launched its sample-based *National Assessment of Educational Progress* (NAEP) in

1969/1970 (Jones 1996; Pellegrino 2014; Johnson 2016); NAEP continues to this day, with some evolution in form and scope (outlined in Section 4). The UK followed closely behind, launching its sample-based *Assessment of Performance Unit* (APU) survey programmes in England, Wales, and Northern Ireland in the late 1970s (Foxman, Hutchison & Bloomfield 1991; Johnson 1989, 2012 Chapter 7, 2016; Newton 2008), modelled to a great extent on NAEP. Unlike NAEP, the APU programme no longer exists, having been rendered defunct in the late 1980s, when the then Conservative government replaced it with a school accountability model (the world's first?). This took the form of a cohort-based *National Curriculum Assessment* (NCA) programme aligned with a newly introduced national curriculum, both of which have been through a series of evolutions since their initial introduction (see, for example, Whetton 2009; Wyse & Torrance 2009; Johnson 2012 Chapter 7, 2016).

Scotland launched its own sample-based *Assessment of Achievement Programme* (AAP) in the mid-1980s (Condie, Robertson & Napuk 2003; Johnson 2016), modelled to some extent on England's APU. The AAP was replaced in the mid-2000s with the more ambitious *Scottish Survey of Achievement* (SSA), which had an additional remit to report attainment by education authority as well as nationally. The SSA was in turn replaced in 2011 by the *Scottish Survey of Literacy and Numeracy* (SSLN), which was politically constrained to adopt a stronger practical skills element to address the demands of the new *Curriculum for Excellence* (Spencer 2013). In 2016, following another international trend, the SSLN was discontinued, and with it sample-based monitoring in general, to make way for a system of online cohort testing. In the event, in response to too many anticipated risks, the new 'standardised testing' programme was never launched in the form planned, leaving system monitoring entirely dependent now on centrally-submitted teacher judgements.

Meanwhile, in the early 1970s, the Republic of Ireland periodically carried out attainment surveys in a variety of subjects (Shiel, Kavanagh & Millar 2014), the year-groups involved in surveys varying each time prior to 2009. In the mid-1980s, the Netherlands launched its primary sector *Periodic National Assessment Programme* (*Peridodiek Peilings Onderzoek,* PPON), which, like NAEP, continues to this day (Scheerens, Ehren, Sleegers & de Leeuw 2012). New Zealand's *National Education Monitoring Programme* (NEMP), which was unique in its focus on reporting performance at the level of individual items and tasks only, for the benefit of the teaching profession rather than policy-makers, was launched in the mid-1990s (Crooks & Flockton 1993), and ran for 15 years, national assessment thereafter relying uniquely on census surveys based on teacher judgement (Flockton 2012). Another country that accumulated large-scale assessment experience in this era was Canada. Activity in this federated country was initially confined to system evaluation in one single region, British Columbia, eventually expanding nationwide. Canada's first national surveys were launched in 1993-4 within the *School Achievement Indicators Program* (SAIP), which was replaced in 2007 by the *Pan-Canadian Assessment Program* (PCAP) – for the final SAIP report (on science) and the most recent PCAP report see, respectively, *Council of Ministers of Education, Canada* (CMEC) (2005) and O'Grady and Houme (2014).

With rare exceptions, assessment in these early programmes took place towards the end of a school year, with the attainment findings often supplemented by questionnaire-based information for contextualisation. France took a different direction when it launched a programme of 'diagnostic' national assessment in the late 1980s (Bonnet 1997; Trosseille & Rocher 2015). This was exhaustive assessment, intended primarily to provide information for school inspectors and receiving teachers about students' strengths and weaknesses as they started a new school year, so that appropriate teaching programmes might be planned for that year. Like NEMP, there was no overt political motivation behind this programme. The principal purpose was to support teaching and learning, and in this way to quality assure system effectiveness. The French government gathered students' test results from a

randomly representative sample of schools for its own analysis purposes, but no analysis results were published. In the mid-2000s this programme was abandoned in favour of accountability programmes, the most recently launched focusing on testing in the early primary school and the beginning and end of the lower secondary school, with surveys online-delivered in the secondary sector (Andreu, Ben Ali & Rocher 2016). A change of government in 2017 saw the reintroduction of 'diagnostic' assessment for the principal benefit of receiving teachers, with testing at the beginning of the entry year into primary education.

Among countries in Eastern Europe, Hungary is one that has not only benefitted from a relatively long history of participation in international survey programmes, but which also has many years of national assessment experience, which began with *TOF-80*, a one-off sample-based survey in various school subjects in Grades 4 and 8 in 1980 (Balázsi 2007). This was followed by a sequence of *Monitor Studies*, which ran for around 20 years from implementation in the mid-1980s, at different grade combinations each time. Hungary today operates annual cohort testing in a subset of year-groups (Grades, 6, 8 and 10), within its *National Assessment of Basic Competencies* (National ABC).

Among developing countries in the southern hemisphere, Chile is notable for its early entry into large-scale assessment. The country participated in one of the IEA's first single-subject surveys, the first science survey of the early 1980s (Johnson 1999), and launched its own domestic system evaluation programme – the *Programa de Evaluación del Rendimiento* (PER) – at around the same time (Ferrer 2006; Gysling 2016). The PER became the *Sistema de Medición de la Calidad de la Educación* (SIMCE), which continues today in the form of a high-stakes cohort-based school accountability programme (Meckes & Carrasco 2010; Gysling 2016).

From its original conception as a research tool, of interest principally to educational researchers, the political significance of the IEA has inevitably and steadily increased. Financial support from the World Bank from the early-1990s helped to put TIMSS and PIRLS onto regular cycles, and continues to ensure the participation of many developing countries in IEA surveys. Survey reports and press releases now give prominence to 'country league tables', that show the relative standing of every participating country in terms of its sample students' performances; see, for example,, the latest available reports on reading (the 2011 survey – Mullis et al. 2012), mathematics (the 2015 survey – Mullis et al. 2016) and science (the 2015 survey – Martin et al. 2016).

2.2 International aid for education reform from the early 1990s

While developed countries around the world benefitted from involvement in the IEA surveys during the Association's first 40 years of activity, few developing countries were so fortunate (Johnson 1999; Kellaghan & Greaney 2001), with participation rates in mathematics and science being generally higher than for reading. No developing country was among the 12 countries that took part in the first IEA mathematics survey of the early 1960s, though three did so in the first science survey – Chile was one of these, as already mentioned, along with India and Thailand. Three developing countries – Nigeria, Swaziland, and Thailand – participated in the second IEA mathematics survey, with 11 taking part in the second science survey.

Several developing countries in Asia, Africa, and Latin America also participated at one stage or another in TIMSS in the mid-1990s, though fewer undertook student attainment surveys than had done so in the previous science survey undertaken 10 years earlier. Several Latin American countries and a number of African countries participated at the planning stage, gaining capacity-building benefit without incurring the prohibitively high costs of full involvement. China, the Dominican Republic, and the Philippines also took part in the

extensive curriculum analysis exercise. But just six developing countries participated fully, carrying out student testing on the scale required: Colombia, Korea, Mexico, Singapore, South Africa, and Thailand (Johnson 1999).

Recognising the value of international system comparison for countries still in the throes of development, the World Bank agreed to fund the participation of around 20 developing countries in what was essentially a late 1990s re-run of TIMSS; sponsored countries included Chile, Chinese Taipei, Hong Kong, Korea, Malaysia, the Philippines, Singapore, South Africa, Thailand, and Tunisia.

Financial considerations clearly dictate the degree of involvement of individual countries in international surveys, while political instability, infrastructural issues and skills shortages can also create problems during survey implementation. The ability of developing countries to participate in international surveys, and to launch and sustain their own domestic survey programmes, has, however, grown over the past two decades, as one important consequence of broadly-based donor-supported education reform initiatives.

2.2.1 UNESCO's Education for All (EFA) initiative

Numerous international donor organisations, along with government and non-governmental aid agencies and private foundations, have been active since the early 1990s in initiating and supporting comprehensive education reform in developing countries around the world, including engaging in in-country capacity building for reform sustainability. The *United Nations Educational, Scientific and Cultural Organization* (UNESCO), in particular, has been a major player in the education reform arena, not least with its *Education for All* (EFA) initiative, which has now spanned more than two decades of activity.

EFA began with the first *World Conference on Education* held in Jomtien, Thailand, in 1990, an event convened jointly by UNESCO, *the United Nations Children's Fund* (UNICEF), the *United Nations Development Programme* (UNDP) and the World Bank. The seminal outcome of this conference was the *World Declaration on Education for All*, with an accompanying *Framework for Action to Meet Basic Learning Needs* (UNESCO 1990). The document articulated the principal goals for the development of quality education provision across the world, including, in particular, providing universal primary education to all children, ensuring gender equity in access and learning opportunities, reducing rates of adult illiteracy, and eliminating poverty. The EFA programme embraced education reform in many countries around the world, and outcomes evaluation was an integral part of every supported country project:

> *… evaluation of outcomes was an important, indeed non-negotiable, component of all country projects under EFA.* (Bohla 2012: 408)

In the early stages of EFA, outcomes evaluation did not include system evaluation through large-scale assessment, but this was to come. In preparation, UNESCO in 1995 established a *Monitoring and Evaluation Section* within its EFA secretariat in Paris, and simultaneously launched its joint UNESCO-UNICEF capacity building *Monitoring Learning Achievement* (MLA) programme (Chinapah 1995).

A decade on from the Jomtien Conference, the *World Education Forum* took place in Dakar, Senegal. The 1,100 participants reviewed progress, reaffirmed a commitment to achieve Education for All by 2015 (UNESCO 2000a), and adopted Frameworks for Action for six world regions (UNESCO 2000b):

1. Sub-Saharan Africa
2. The Americas
3. The Arab States

4. Asia and the Pacific
5. Europe and North America
6. E-9 countries (high population).

UNESCO was entrusted with overall responsibility for the future coordination of international players and for sustaining the global momentum.

When 2015 arrived, UNESCO and partners took stock of progress at the World Education Forum in Incheon, Korea. A major outcome was a reaffirmation of:

> *… the vision of the worldwide movement for Education for All initiated in Jomtien in 1990 and reiterated in Dakar in 2000 — the most important commitment to education in recent decades and which has helped drive significant progress in education.* (UNESCO 2015: 1)

New goals were agreed for Education 2030 in the *Incheon Declaration* (UNESCO 2015), to further build on the new EFA vision, encapsulated in the sustainable development goal: "Ensure inclusive and equitable quality education and promote life-long learning opportunities for all" (UNESCO 2015: 1). An agreed element in implementation plans foresees an expansion of national assessment activity to serve both national system evaluation and EFA reform evaluation:

> *We resolve to develop comprehensive national monitoring and evaluation systems in order to generate sound evidence for policy formulation and the management of education systems as well as to ensure accountability. (UNESCO 2015: 4)*

2.2.2 SACMEQ regional collaborations: Southern and Eastern Africa

The Southern and Eastern Africa Consortium for Monitoring Educational Quality (**SACMEQ**), a network of 15 ministries of education in Anglophone countries, essentially originated in the EFA movement. Formally launched in late 1995, with long-term support provided by UNESCO's *International Institute for Educational Planning* (IIEP) and the Government of the Netherlands, Consortium membership initially comprised seven African governments (Kenya, Malawi, Mauritius, Namibia, Tanzania (Mainland and Zanzibar), Zambia, and Zimbabwe), others joining later when budgets permitted (Botswana, Lesotho, Mozambique, Seychelles, South Africa, Swaziland, and Uganda). SACMEQ's mission has been to expand opportunities for educational planners to gain the technical skills required to monitor and evaluate the quality of their education systems, and to generate information that can be used by decision-makers to plan and improve the quality of education.

With World Bank support, the Consortium has so far undertaken three large-scale, cross-national studies of the quality of education, all of which have focused on Grade 6 learners (end of primary school): SACMEQ I (1995-1999, reading) with 7 countries; SACMEQ II (2000-2004, reading and mathematics) with 14 countries; and SACMEQ III (2006-2010, reading, mathematics, and HIV and AIDS knowledge) with all of the current 15 countries. Surveys included the administration of tests to samples of Grade 6 learners, and of background questionnaires to teachers and school principals.

Once the surveys under SACMEQ III were completed, attainment results and 'conditions of learning' information were available for over-time and cross-border comparison (see, for example, Makuwa 2010 and Spaull 2011). SACMEQ IV was scheduled to take place in 2012-2014, with the 15 existing SACMEQ members plus Angola as an observer and potential future participant, while in February 2015 the SACMEQ Coordinating Centre moved from IIEP to the University of Gaborone in Botswana to be closer to the National Research Coordinators. Reports from SACMEQ IV are awaited.

With World Bank or UNICEF funding, and after having benefited from the capacity building support for sample-based large-scale assessment that SACMEQ participation had offered them, many SACMEQ countries eventually launched their own domestic survey programmes – most based, interestingly, on cohort assessment (!). Examples include Lesotho, Malawi, South Africa, Uganda, and Zimbabwe; Section 4 provides some details.

2.2.3 PASEC regional collaborations: Francophone African countries

The *Conference of Ministers of Education of French-speaking Africa*, or CONFEMEN (*Conférence des Ministres de l'Éducation des États et Gouvernements de la Francophonie*), was originally set up in 1960 by 15 African states, before EFA was launched. The organisation expanded steadily both in numbers of members and range of activities over subsequent decades. Today CONFEMEN is the principal forum for the exchange of information on education policy among the governments of 44 French-speaking countries worldwide. Among the many initiatives sponsored by the organisation is the *Programme of Analysis of Educational Systems* (*Programme d'Analyse des Systèmes Éducatifs*) or **PASEC**, introduced at the 43rd CONFEMEN ministerial summit, held in Djibouti in 1991. The original aim of PASEC was to reinforce francophone solidarity, and to support informed choice of educational strategies.

The first PASEC-inspired national attainment surveys were carried out during the 1993-4 school year in Congo, Djibouti, and Mali, followed by Senegal and the Central African Republic in the following year, under the supervision of the Universities of Mons, Montréal and Laval, and the French International Centre for Studies in Education (*Centre International d'Études Pédagogiques*). Following these initial surveys, the STP (*Secrétariat Technique Permanent*) assumed control until 2013. Surveys continued along similar lines during the period of direct STP management, with single-country surveys, sometimes repeated, in several African countries and also in Lebanon. In 2011-12 the PASEC surveys were extended beyond African borders into Laos, Cambodia (PASEC 2014a), and Vietnam (PASEC 2014b). Gradually, during this period, a variety of external organisations collaborated in various ways: the World Bank, with a first intervention in 2004 when it commissioned a study of primary teachers in nine countries, the French Ministry of Education, UNESCO (regional office for education in Africa), UNICEF, and the IEA.

In a new venture, the 54th ministerial meeting in Chad decided in late 2012 to move to a system of international comparisons of student attainment in French and mathematics at the beginning and end of primary education. 'PASEC 2014' was designed in 2013 and implemented in 2014 in ten countries of sub-Saharan Africa (PASEC 2015): Benin, Burkina Faso, Burundi, Cameroon, Chad, Congo, Ivory Coast, Niger, Senegal, and Togo. Additional finance was supplied by the World Bank, the French Development Agency (AFD), and the Swiss Department for Development and Cooperation. Full details of the initiative and its findings are given in PASEC (2015). Another round of cross-border surveys is scheduled for 2019, with the same focus as PASEC 2014, (i.e., testing at the beginning and end of primary education). PASEC 2019 is expected to involve 15 African countries: all 10 of those that took part in PASEC 2014, with new entrants the Democratic Republic of Congo, Gabon, Madagascar, Mali, and Mauritius.

2.2.4 LLECE regional activity: Latin America and the Caribbean

Education reform throughout Latin America and the Caribbean has been heavily supported by a number of international organisations, among which the *United States Agency for International Development* (USAID), the IEA, the World Bank, the Inter-American Development Bank, the Tinker Foundation, and the GE Foundation. These organisations financially support the influential *Partnership for Educational Revitalization in the Americas* (PREAL), which was established in 1995 by the *Inter-American Dialogue* in Washington, D.C. and the *Corporation for Development Research* in Santiago as a multi-year initiative to

build a broad and active constituency for education reform in many Latin American countries. In 2002, PREAL's *Working Group on Assessment and Standards* began carrying out studies on how the quality of education was being measured in Latin America, including monitoring progress in how well expectations for learning were being articulated. Findings up to the mid-2000s are usefully summarised and evaluated by Ferrer (2006), and updated by Ferrer and Fiszbein (2015).

UNESCO works in coordination with the countries of the region through its *Latin American Laboratory for Assessment of the Quality of Education* (LLECE), a network of national education quality assessment directors across Latin America and the Caribbean. LLECE organises comparative studies aiming to measure the quality of education in the region. *The First Regional Comparative and Explanatory Study* (PERCE, by its Spanish acronym) was carried out in 1997 in 13 Latin American countries. In 2006 the Second Study (SERCE) saw the participation of 16 Latin American countries plus one Mexican state. The Third Study (TERCE) was implemented in 2013 in 15 Latin American countries and the one Mexican state. SERCE and TERCE were directly comparable studies allowing relative attainment and over-time attainment change to be recorded for those participants that took part in both surveys: Argentina, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Nicaragua, Paraguay, Peru, Panama, Uruguay, and the Mexican state of Nuevo Leon (ACER 2014).

2.2.5 The Post-Socialist Education Reform Package

The EFA initiative had its origins in a desire to help improve educational opportunities and quality for populations in developing countries around the world that remained in states of relative poverty, afflicted by natural disasters and conflict. Its principal focus was the primary sector of education, and universal access and equity were among its earliest high-priority goals.

It happened, though, that just as the EFA movement was unfolding the Soviet Union was being dismantled, leaving major regions of the world, and the newly independent countries within them, with different education reform needs. Primary education here was already universal, or nearly so. This was therefore not a concern. What was needed, according to the governments in the countries concerned, and the international community that came to their aid, was assistance with system modernisation, in both primary and secondary (and tertiary) education. Access and equity were less of an issue than system quality – in the sense of systems meeting the educational needs of the modern world. In response, through the 1990s and 2000s, numerous international donors, aid agencies, NGOs and private foundations jointly provided what became known as the *Post-Socialist Education Reform Package*. In an authoritative and critical account of the external impact of international interventions through this period in the Caucasus and Central Asia, Silova and Khamsi (2008) note that:

> *External influences came most visibly in the form of foreign aid, which boomed in the early 1990s, and then escalated further at the start of the millennium.*
>
> (Silova & Khamsi 2008: 4)

Among the largest agencies at play in the Caucasus and Central Asia from the outset were the UN, the World Bank, the *International Monetary Fund* (IMF), the *European Bank for Reconstruction and Development* (EBRD), and the *Asian Development Bank* (ADB). By the beginning of 1992, these organisations had already conducted assessment missions in several countries, including Armenia, Azerbaijan, Georgia, Kazakhstan, Tajikistan, Turkmenistan, and Uzbekistan, and were preparing for reform support (Silova & Khamsi 2008). The *European Union* (EU), the World Bank, UNESCO, and other organisations, including UK awarding bodies, provided financial support and expert assistance to several

other Central and Eastern European governments during the same period – Romania, Hungary, Lithuania, Poland, Slovenia, and the Russian Federation among many others – in planning for and implementing education system reform. Reforms embraced a liberalisation of previously tightly controlled policies on curriculum, textbooks and teacher education, along with the introduction of new forms of student assessment, a tightening of control over school leaving examinations, and, in some cases, preparation for national assessment.

West and Crighton (1999) provide a comprehensive overview of reform ambitions, with a particular focus on the reform of school-leaving examinations. Bakker (1999) describes a Russian-Dutch project aimed at standardising school-leaving examinations across the Russian Federation, the Dutch Educational Research Institute (CITO) taking a central support role, as it continues to do throughout Central and Eastern Europe. Extension of education system reform to system monitoring through large-scale student assessment followed in some countries: Bialecki, Johnson and Thorpe 2002 offer a rare country case study, with a focus on Poland. Progress in national assessment was soon to accelerate through the activities of the OECD and in particular its PISA programme.

2.3 The influence of the OECD's PISA from 2000 onwards

In 2000, the world witnessed the launch of what arguably continues to be the greatest influencer of all on educational politics internationally. This is the OECD's **PISA,** whose initial primary purpose was to provide the OECD with comparative information about the output of its member countries' educational systems, in terms of student attainment at age 15 (the end of compulsory schooling in many countries at the time). Information about every other aspect of educational provision in OECD countries – structure, input, and process – had been, and still is, regularly documented in the OECD's annual *Education at a Glance* reports, but no complementary information about outcomes was available to complete the picture.

As mentioned earlier, until the mid-1990s the IEA's surveys were sporadic, they involved a different subset of countries on each occasion, and they focused on age groups within and not at the end of compulsory schooling. IEA survey findings could not therefore meet the OECD's particular needs. Where countries had domestic attainment survey programmes in place, these took different forms with different methods of reporting, and did not in every case assess student attainment at age 15. Given these shortcomings, neither the existing international survey programmes nor any active national assessment programmes had the potential to provide the OECD with the regular and comparable outcomes information that it newly sought. And so PISA was born.

PISA surveys take place on a 3-year cycle. To avoid the interpretation issues that the IEA faced when it based its surveys on commonality in national curricula, the original PISA assessment frameworks 'transcend' national curricula by assessing 'skills for life' (OECD 2009: 14):

Reading literacy: An individual's capacity to understand, use, reflect on and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society.

Mathematical literacy: An individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen.

Scientific literacy: An individual's scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues, understanding of the

characteristic features of science as a form of human knowledge and enquiry, awareness of how science and technology shape our material, intellectual, and cultural environments, and willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen.

Every survey incorporates assessment of all three literacies. On each occasion, one type of literacy assumes 'major domain' status, and consumes two-thirds or more of the testing space, while the other two literacies are 'minor domains'. Each type of literacy is assessed as a major domain in every third survey, (i.e., every nine years). The latest survey to date is that of 2015, which for the second time has included the assessment of problem-solving, and in which for the first time assessment in most participating countries was computer-based (OECD 2016a, 2016b).

The first PISA survey report (OECD 2001), and in particular the country 'league table' presented within it, caused more than a flurry of interest among politicians and policy-makers around the world. It offered satisfaction to those whose countries appeared towards the top of the league table (notably Finland among European countries, along with several Asian countries, including Japan, and Korea), inspired resignation in those whose countries were placed towards the bottom, and surprised and dismayed some of those whose countries performed much less well than expected when compared with neighbouring countries with similar states of socioeconomic development (other Scandinavian countries, for example,, in the case of Finland). There was indeed 'PISA shock' in some cases. Long-established system structures, which until PISA had been assumed to function satisfactorily, were suddenly under review. Germany, Norway, and Switzerland are well-known examples (Baird et al. 2011, 2016), with Wales joining them when its results, reported after its first survey participation in 2006, proved to be significantly below those of the other constituent countries of the UK in mathematics and reading (Bradshaw et al. 2007).

Inevitably, league table interest has been maintained as further PISA surveys have been reported, with headline summary statements such as the following from a 'focus report' on the 2012 survey:

> *Shanghai-China has the highest scores in **mathematics**, with a mean score of 613 points – 119 points, or the equivalent of nearly three years of schooling, above the OECD average. Singapore, Hong Kong-China, Chinese Taipei, Korea, Macao-China, Japan, Liechtenstein, Switzerland and the Netherlands, in descending order of their scores, round out the top ten performers in mathematics.* (OECD 2014: 4, original emphasis)

An important outcome of the 'PISA phenomenon' has been a rapid growth of enthusiasm internationally for the development of national assessment programmes, to provide national student attainment data, along with contextualising information about learning environments, that those responsible for ensuring the effectiveness of their national education systems more urgently demand. Germany is a much cited example, with its federated politics and segregated academic/vocational education system – a country which, like many others, had not pre-PISA had national student attainment data available for its own system evaluation purposes. Switzerland is another, as are Australia, Brazil, Canada, and many, many more. With some inevitable exceptions, the numerous developing countries around the world that received capacity building support through the long-running EFA initiative were to a greater or lesser extent prepared to address their new ambitions to benefit from their own domestic system evaluation programmes.

3. Design choices for programme development and issues arising

If national assessment programmes are to function successfully, meeting their intended purposes adequately, then they must be designed to provide information that is dependable (valid and technically reliable) and unbiased, based on data that can be gathered as economically as possible and with minimal disruption in schools (Johnson 2016). Designing a national assessment programme to satisfy these demands can be a challenging exercise, involving a combination of political, professional, and technical decisions about a number of different features. These include:

- curriculum focus
- age-stages for assessment
- scale of assessment
- survey cycle and timing
- modes of assessment
- underpinning measurement model
- attainment reporting
- dissemination strategy
- programme management.

3.1 Curriculum focus

A critical, and usually uniquely political, decision for national assessment concerns the curriculum focus of the new programme. Should this be a focus on 'key' curriculum subjects, and be confined to a small subset of these, with domain and subdomain reporting as appropriate? The subset would today, post-PISA, typically comprise the language of instruction and mathematics (in the primary sector 'literacy' and 'numeracy'), with science also occasionally featuring. Arguments for the 'key subjects' approach include the acknowledged special importance of those subjects in terms of providing students with essential knowledge and skills for learning and further learning, across the curriculum and throughout life. Financial, logistic, and technical arguments would also favour a subset of subjects with little or no requirement for practical assessment to be included (lower cost, easier implementation, higher technical quality – though possibly with lower assessment validity).

Or should programme coverage be broader than 'key subjects', and cover other curriculum subjects as well, such as history, geography, music, art, as the IEA achieved in its early cross-border surveys, as NEMP in New Zealand succeeded in doing during its lifetime, as the APU in the UK and the SSA in Scotland did briefly with Technology and Social Subjects Enquiry Skills, respectively, and as NAEP in the US does today? Narrowing a programme's sphere of interest can serve to 'devalue' those subjects not overtly acknowledged as 'key'. There was, for example, concern expressed for history when the UK's APU was first launched, with its focus on language, mathematics and science. Subject exclusion from national assessment can also result in general curriculum distortion, for example, with schools modifying instructional time allocations to different subjects accordingly; when science was dropped from the NCA in England the time devoted to this subject area in schools apparently diminished.

Given such curriculum issues, should the programme be deliberately unconcerned with the specifics of the school curriculum, and, adopting the PISA approach, aim to assess intellectual skills whose development is considered to be attributable to schooling in general? The risk here is that it would be difficult to identify any curriculum-related policy initiatives that might be needed to improve the situation in some attainment areas. Curriculum-focused national assessment implies that a national curriculum exists – without a national curriculum, differences in population attainment in different regions of the country,

or in different types of school, could be difficult to interpret for the purpose of identifying relevant policy implications. The APU in the UK, particularly the science monitoring programme, operated at a time when no national curriculum existed in the subjects assessed covering primary and lower secondary schooling. The APU science focus on 'process skills' was an acknowledgement of this, and an intended solution. But it was quickly discovered that 'process' is inseparable from 'content' in some areas (most obviously 'applying science concepts', but also 'observing' and 'planning investigations'). This rendered interpretation of findings problematic, especially at age 15, by which time students were studying one, or more than one, science subject, or none at all.

This requirement for attainment results to be clearly interpretable in terms of guiding any follow-on policy intervention in the system means that enthusiasm for the implementation of system evaluation through national assessment must be preceded by the implementation of a national curriculum (at least for the subjects to be assessed) where none already exists, and where the introduction of a national curriculum is a practical possibility. In the absence of a national curriculum when the APU was operating, the assessment programme became seen by many in the education field in the UK as a curriculum driver – the phrase 'assessment-led curriculum' carried negative connotations at the time.

In response to PISA, and in preparation for national assessment, some countries have already taken steps to develop and implement national or regional curricula to replace the variety of more localised curricula that might have existed before. Switzerland, which faces a particular challenge in this regard, given its multilingual heritage, is just one example. This country is in the process of a compromise, replacing numerous cantonal curricula with a common curriculum for each linguistic region (French-speaking, German-speaking, and Italian-speaking) in readiness for the implementation of a 'national' assessment system (SKBF/CSRE 2014). Moves are underway, too, to newly introduce national curricula, or to replace existing national curricula, in developing countries. For example, as mentioned in the brief case study of Rwanda in Section 4, a 'competence-based' *2013 Harmonised Curriculum Framework for the East African Community* has recently been developed with USAID support. For another example, several countries in southeast Asia, collaborating under the guidance of the *Southeast Asian Ministers of Education Organization* (SEAMEO), with support from the *Australian Council for Educational Research* (ACER), have been working towards common core curricula for literacy, numeracy and global citizenship in preparation for cross-border use of the new *Southeast Asia Primary Learning Metric* (SEA-PLM) (ACER 2016a).

3.2 Age-stage choices

It is generally also national politicians who decide on which age groups or school stages should be the focus of a new or redesigned national assessment programme, although their decisions might be modified through circumstance before or after a programme is launched. Popular choices are the end of primary schooling, the case in most developed and developing countries in which national assessment already features, and the end of compulsory schooling. But there are other options available, and in operation.

In the APU programme, for example,, attainment monitoring in science was added at age 13, since this was the age after which optional choices in science, including no further science study at all, came into play, with predictable interpretational problems for survey reporting at age 15, as noted earlier. In the NCA in England, *Key Stage 1 and 2* (KS1 and KS2) are the beginning and end of primary school (for most pupils), with *Key Stage 3* (KS3) the lower secondary school just before study programmes for national qualifications begin (though testing at this stage has been dropped). In France's most recently introduced national assessment programme (Andreu, Ben Ali & Rocher 2016), the initial intention was to assess students in the lower primary school, at the end of primary school and at the end

of secondary school, all recognised as key stages in the recently revised national curriculum (the *socle commun* – MEN 2015; Jeantheau & Johnson 2016). It was also planned to computer-deliver tests as soon as possible, following the PISA model. However, piloting confirmed that primary schools remained less well-equipped for online testing than secondary schools, so for pragmatic reasons the choice for testing at the end of primary school was changed to testing in the first year of secondary school, and this is the stage at which the first survey in the new programme was conducted (Andreu, Ben Ali & Rocher 2016).

A fairly recent international trend, in developed as well as developing countries, is for national assessment to feature at *every* stage, or at several sequential stages, in schooling. The rationale is that this strategy in principle facilitates the monitoring of age-related progression, and, if cohort testing is also implemented, permits school-level reporting. Several countries, large and small, in different parts of the world, have adopted this strategy, including, for example, Australia, Chile, England, Hungary, South Africa, and Qatar. Scotland is one of the latest countries that planned to join this group, with national cohort testing scheduled for launch in 2017 at the beginning, middle, and end of primary schooling and the middle of lower secondary schooling; the new programme was in fact abandoned before launch in response to a concerning risk assessment.

The preschool sector is not immune to national assessment activity, with some countries (e.g., England) being keen to establish a 'baseline' performance profile to serve as a reference for more formal school value-added assessment in the primary and secondary sectors.

3.3 Survey cycle and timing

Few education professionals would expect population attainment change to be apparent after a mere year or two. The relationship between change in curriculum, instructional approaches, learning environments, and so on, and attainment change, must logically take time to have an effect. Thus, long before PISA was launched, with its 9-year cycle for major domains, many countries with national assessment programmes had adopted non-annual cycles, at least for the same subject areas or school stages. This has changed and continues to change.

England's NCA assesses reading and numeracy at a limited number of key school stages, but annually. Several other countries have also adopted this general model at some point, but with an increasingly popular pattern of annual assessment of literacy and numeracy in several *consecutive* school stages (e.g., Brazil, Chile, and South Africa).

France is one of the few countries with a different pattern of operation. The new assessment programme will assess two curriculum areas (French and mathematics/science) in each annual survey, but each time at just one of three stages, with each stage assessed on a 3-year cycle.

As far as the timing of surveys within the school year is concerned, this is most commonly towards the end of a school year, particularly where the outcomes of schooling at the end of a single school stage or multi-stage period are concerned. While this might be considered the ideal, factors of school life can obviate this possibility: educational outings in the primary school, for example, or preparation for external examinations in the secondary school. Where progress over a phase of schooling is of interest, for example, through the duration of lower secondary schooling, then one survey might take place after the start of the initial school year of the phase (e.g., in November, when students can be assumed to have settled in) and the other towards the end of the final school year (typically in May), as in France.

3.4 Mode(s) of assessment

The mode(s) of assessment adopted for a national assessment programme will depend on a number of factors, including the subject to be assessed, the stage or age group in focus, and the state of economic development of the country.

Where children are old enough for the majority to be assumed to be able to read and write, testing could be entirely paper-based, with class teachers or visiting test administrators organising and supervising test sessions, often with the freedom to read questions to individual pupils or the entire group to ensure understanding (though without a warrant to help with answers – an important distinction that survey managers would need to trust class teachers to understand and apply faithfully). Practical assessment tasks are an added option in subjects like numeracy and science, though while their inclusion should increase assessment validity in terms of curriculum coverage in that subject area it will also incur higher cost and could reduce overall assessment reliability.

Marking might be carried out in the schools by class teachers. This would be the cheapest option, but one which again requires a high degree of trust on the part of survey managers. Alternatively, itinerant test administrators could be given this role, or scripts (and possibly videos or artefacts in the case of performance assessments) might be securely transported to central locations for this purpose.

Until very recently, testing in literacy and numeracy has been paper-based in most countries, developed and developing, and continues to be so in many. Where technology availability in schools allows, tests might be delivered electronically, on CD-ROM or using web-based applications. Accelerated through the encouragement/pressure of the OECD, with its move to electronic testing within PISA, many countries around the world have indeed moved from paper-based to electronically-delivered testing in secondary schools; France is but one example already mentioned, while Scotland used a mixture of paper-based and online testing in the SSLN (every student subject to both modes of delivery for their test batteries), as did Kazakhstan in its EALA (each student experiencing one or other form of delivery, depending on computer availability in their schools). Some eastern European countries with very recently introduced national assessment programmes adopted electronic delivery from the start.

Where children are too young to respond independently to paper-based or electronically delivered assessment materials, for example, in the pre-primary sector or in the lower grades of primary school, then assessment can rely on class teacher judgement, based on observation of behaviour and development over a period of time. Alternatively, assessment can be organised through one-to-one interaction with teachers or trained administrators (the Early Grade Reading Assessment (EGRA) and the Early Grade Mathematics Assessment (EGMA) are examples). National assessment can also rely on teacher judgement later in schooling, wholly or in part. England's NCA is an example; this initially relied entirely on teacher assessment, then on paper-based testing, and now functions with a mixture of both forms of assessment – tests for reading comprehension and numeracy, teacher assessment for writing.

3.5 Scale of assessment

As to the scale of testing, this can take any one of a number of different patterns, with or without student, class or school sampling. The most common strategies are to test or otherwise assess:

- a sample of students from across a stage or age group in a sample of schools
- a sample of students from across a stage or age group in every school
- all students in one or more classes in a sample of schools
- all students in a stage or age group in a sample of schools

- all students in a stage in every school (cohort testing).

Sample-based surveys are in principle more cost-efficient than cohort testing could ever be. However, it is critical that the drawn student sample is appropriately representative (or is deliberately non-representative) of the national population at the age/stage concerned, and in an ideal world the achieved sample should match intentions. A nationally representative student sample will be one where the proportion of students in the sample of any given type (e.g., girls in small schools in a particular region) matches the proportion in the population as a whole. A proportionate stratified sampling scheme will produce such a sample. A deliberately non-representative student sample will be one where particular subgroups will be under-sampled or over-sampled relative to their population presence, in order to have sufficiently large subgroup samples available for dependable comparative attainment data to be produced (for gender comparisons, regional comparisons, socioeconomic comparisons, and so on). A disproportionate stratified sampling scheme will be employed here.

Where samples are disproportionate by design, or where samples are intended to be directly representative of their populations but are not, for example, because of student absences or withdrawals, teacher strikes, school crises, and the like, then data weighting will be employed to redress the imbalances when national attainment estimates are produced. Note, though, that redressing observable imbalances through data weighting will not necessarily simultaneously address any potential biases that are not observable. Where achieved samples fall short of intentions, in terms of composition or size, then the dependability of national attainment estimates and of subgroup comparisons will be jeopardised.

One design option not so far mentioned is the choice between strictly cross-sectional surveys and pseudo-longitudinal or genuinely longitudinal surveys. In a sample-based programme comprising cross-sectional surveys, whenever a school stage is to be assessed the sample of students drawn for testing, or the entire cohort, will be a new one, and the attainment results will apply to that stage in that survey year only. Even then, where more than one stage is assessed in each survey, Grades 4 and 6, say, some educated inferences about age-related learning progression should be possible. In pseudo-longitudinal surveys, where different stages are assessed in each survey, and where the number of school years between those stages matches the survey cycle (e.g., Grade 4 in one survey and Grade 6 in the next in a programme whose surveys are conducted on a 2-year cycle), then, even though the sampled students assessed might be different in each stage in each survey, some inferences might be drawn about age-related learning progression over the intervening period of schooling, and hypothesised to reflect the impact of policy initiatives taken in the meantime. A genuine longitudinal survey would be one where the same students assessed in the survey at the earlier stage are re-assessed again in the follow-on survey at the higher stage. Problems here, however, would include a potential threat to student anonymity (this would be an issue in countries where anonymity is guaranteed), and sample size attrition (cohort assessment would solve this, but at much greater cost).

3.6 Underpinning measurement model

Organising and implementing a one-off attainment survey in any subject domain can never be without challenges, in terms particularly of logistics. But the real challenges arrive when a single survey is followed by others in the same subject domain, with the explicit intention to monitor ‘standards’ over time.

3.6.1 Common items

One possible strategy for over-time monitoring is to use the same set of items, tasks, or tests each time a subject survey is carried out, in the same form, administered in the same

way and with unaltered mark schemes. While appealingly simple, this practice is not advised for a number of reasons. Firstly, unless the time between surveys is extremely large – several years – then there will be a risk of test, or at least item, exposure jeopardising the validity of the attainment results of the later surveys. Whether a single test is administered in a subject survey, or a number of alternative tests are used, if these, or at least a subset of exemplary items, are not released after use then stakeholders will have difficulty interpreting, or even believing, the reported survey findings, and will not have clear guidance on future action. Teachers' professional development will also be difficult to organise without jeopardising the validity of the results of future surveys. Finally, over a long period of time the test(s) can also become less and less relevant to a constantly evolving curriculum, and in consequence lose validity and policy value.

3.6.2 Domain sampling and item response modelling

Superior strategies for 'test' creation for use in large-scale attainment surveys are domain sampling, a rarely used approach, but one with high potential, and the application of item response theory (IRT), also known as item response modelling, the almost universal choice for national assessment post-PISA. In principle, both approaches require the creation of a large pool of relevant test items before programme launch, although this rarely happens. The item pool, or item bank, represents the subject domain(s) of interest, along with relevant subdomains, its composition being defined, constrained and controlled by a 'pool specification', taking into account item types, knowledge/skills assessed, mode of delivery, and so on, perhaps with some empirical properties included.

In domain sampling, the whole set of items to be administered in a survey would be drawn from the pool using stratified random sampling (Johnson 1989, Chapter 3), following a given survey specification, and then subdivided into a series of interchangeable tests. England's APU used this approach for its science assessment and it was successfully used also for mathematics assessment in Scotland's SSA and SSLN. The national student sample, or entire student cohort in cohort assessment, would also be randomly subdivided, into equivalent subsamples in terms of composition (gender, school type, and so on). Using an appropriate strategy tests are then randomly administered to student subsamples within the survey. This is matrix sampling. In an IRT application, for which items would first need to have been calibrated for difficulty in large-scale pretesting, items would again be drawn from the pool to create subsets for matrix administration to student subsamples, with in-built item overlap across student 'test packages'. Most countries around the world that operate national assessment programmes have, with international support, adopted the IRT approach, following the PISA model. But this model is not without its own issues.

An important feature in IRT applications is that there is no explicit acknowledgement of item sampling, so that any contributions to attainment estimation error due to this sampling element are not accounted for. There are issues, too, surrounding some strong assumptions about item behaviour, or, rather, about the effect, or lack of effect, of topic exposure on relative item difficulty; the Rasch model, in particular, assumes that relative item difficulty is invariant across student subgroups, whatever their prior learning experience, and over time. Where this assumption is tested empirically, 'differentially functioning items' will be excluded, raising questions about the validity of those remaining as a faithful representation of the subject domain concerned (in literacy and numeracy this potential problem is less serious than it has been found to be in science). This assumption of the invariance of relative item difficulty is a fundamental one, and while it can be evaluated for subgroups at a single point in time, it is not possible to evaluate invariance in the same way over a future period of time, other than retrospectively. Thus it is that IRT-based programmes are often well into their monitoring role before attainment anomalies begin to emerge and attainment trend data is disrupted (see OECD 2016b, Annex A, for relevant comment with regard to PISA).

3.7 Attainment reporting

There are a number of ways in which attainment results can be presented for stakeholder consumption. Item-level results, for example,, will be given as facility values, 'percent correct', in the case of binary-scored items, and can be offered as mean scores or frequency distributions (over the item mark scale), or even as 'mastered/not mastered' percentages, in the case of non-binary items and multi-step tasks.

When it comes to reporting domain attainment summatively, again there are choices. Achievement might be reported in terms of mean scores or mean percentage scores, for a single test, if common across the sample or cohort, or averaged over multiple interchangeable tests in a matrix sampling context; alternatively, mean scores across relevant sets of individual items are possible, where tests (or 'item packages') comprise items relating to different aspects of curriculum. In place of mean raw scores or mean percentage scores, reporting can be in terms of mean scale scores (as in IRT applications), with reference to a scale with a pre-determined, essentially arbitrary, mean and standard deviation (such as 500 and 100, respectively).

Mean scores of any type, however, have proven difficult for stakeholders to make meaning from in terms of general 'standards' of achievement and in terms of what they say (nothing in fact) about the interpretable *nature* of that achievement. For example, what does it mean for a politician or a teacher to be told that the estimated mean score of the nation's 10-year-olds in reading is 70%, or 50%, or 55% (or 495, 502 or 525)? Whichever figure might be reported, how can anyone judge whether this is a 'good' result, an 'adequate' result, or a 'poor' result, and, if the latter, in what ways, and with what policy initiatives might the problem be addressed? Mean scores can become more meaningful when compared with others, for example, in subgroup comparisons (boys versus girls) or over time (a fluctuating, underlying 'stable', attainment picture, or rising or falling trends). The APU was reported in this way in its short lifespan, and was reasonably well accepted by all stakeholder groups initially. But once the population attainment level was known, along with the relative standing of key student subgroups, and the picture was essentially unchanged from one survey to another, stakeholder interest waned. Politicians, in particular, demanded more detailed information, that they could better interpret and use in policy formulation. The question now was: '*but what do students know* and *what can they do*?'. The programme had never been designed to provide this kind of information, and was given no time to remodel itself to be able to do so.

Politicians in England had also become more and more focused on the issue of school accountability, and were frustrated by the fact that the APU could not, again by design, provide school-level attainment information.

When mean scores alone are not enough for policy purposes, then an alternative is to group students into 'performance bands' on the basis of their achievement scores, and to report the proportions of students in each band. When bands have evocative labels attached, such as 'basic', 'proficient', 'advanced', and when short verbal descriptions of the 'meaning' of these labels in terms of the knowledge and skills acquired by those students classified into one or other group are offered, then survey results take on meaning and become 'user friendly' for all stakeholder groups.

But assigning students to performance bands, either directly on the basis of teacher judgement or through the application of threshold test scores, is not an exercise that is without challenge and risk. The risks have to do with validity and reliability, and stability. Despite the lower popularity of mean scores, the Statistics Commission (2005), on the basis

of evidence from a review of stakeholder understanding of survey attainment results, recommended their use:

> *Frequently statistics are quoted in terms of the percentage achieving some fixed threshold and then changes in these percentages are regarded as valid and reliable measures of overall improvement for different groups. However, simple models show that these measures can be very misleading and it is better to base statistics on averages of performance over a whole cohort (e.g., average point scores), which are much less sensitive to changes by small groups at some arbitrary boundary.*
>
> (Statistics Commission 2005: 58)

3.8 Dissemination strategy

Ideally, the members of every stakeholder group should have rapid access to easily digestible survey results, particularly where survey findings will be relevant to their interests and could determine productive future action on their part. It can be assumed that national and local politicians and policy-makers, along with schools inspectors, will have such access. Teacher trainers in universities and colleges, too, could receive salient findings in some form.

Schools can, and should, be included in a dissemination strategy, so that teachers can be kept informed about the state of the nation in terms of student attainment, particularly strengths and weaknesses, and learning environments. This stakeholder group, however, has perhaps proved the most problematic to reach. Some developed countries have in the past distributed full paper-based survey reports to schools, on the assumption that head teachers, to whom reports were typically addressed in the first instance, would draw the attention of their staff to them; experience, however, suggested that teachers did not always receive the documentation, or, if they did, they did not necessarily consult it, for whatever reason. Themed reports, 'teaching and learning points', and other more targeted information devices have also been produced, in paper-based and electronic form, but have been little used apparently. Yet teacher buy-in to national assessment is essential if the quality of a programme is to be assured, and buy-in can only be improved through a broader and better understanding of what attainment surveys are about and what their value to the system and to the players within it can be – this implies more effective dissemination within the teaching profession.

In Scotland, the student sampling strategy adopted for the SSLN was unique, in that there was no sampling of schools or classes. All eligible schools were invited to take part in surveys, but with very small numbers of students then randomly selected for assessment in each participating school (2 in primary schools, and a maximum of 12 in secondary schools). There were a number of reasons for the change from the usual pattern of 2-stage school-student sampling. Firstly, it was to ensure that the SSLN, unlike the SSA, could not under any circumstances be expected to provide attainment data for use in school comparisons, thus effectively neutralising previous pressure from the schools inspectorate for just this kind of information. Secondly, it was expected to facilitate non-disruptive practical assessment in survey schools, where practical assessment had now become a requirement, rather than an option, to align with the Curriculum for Excellence. Thirdly, and this is the relevant point here, it was in the hope of increasing SSLN awareness among all the nation's teachers, rather than just those in the previous samples of participating schools, so that survey findings might have a better chance than before of reaching the whole profession on a regular basis.

Television coverage represents another potential dissemination strategy, one that was used in Africa to highlight some of the findings from the PASEC 2014 series of national surveys in francophone countries.

3.9 Programme management

A final point worth noting briefly here is that national assessment programmes need to be managed, and top-level decisions must be made in regard to this. When launch of a new programme is contemplated, an organisation, or group of partner organisations, must be identified and commissioned to manage or co-manage programme planning, implementation and reporting. Budgets, too, must be agreed.

Programmes can be operated from within the government itself, if sufficient staff with the necessary administrative, professional and technical experience and expertise pre-exist, or can be hired. Another arrangement can take the form of the government having overall control, but with quasi-governmental organisations and/or universities working in partnership, perhaps one handling school liaison and testing, another taking responsibility for test material development and exemplification, a third undertaking primary and secondary data analysis, and a fourth addressing report writing and dissemination.

An alternative model is for governments to commission external testing agencies to undertake all or some aspects of programme development and operation, with or without donor support.

4. Programme models: brief case studies from around the world

The following overviews of activity in different world regions over the past two decades will illustrate the pace and scale of the expansion of national assessment internationally, and the variety of programme models adopted. The selection of brief country case studies offered should convey an idea of history and evolution, and also of the kinds of problems that have been met in practice.

4.1 North America

4.1.1 The US

The *National Assessment of Educational Progress* (NAEP), the first and the longest lasting system evaluation programme, was launched in the US in the early 1970s (for a comprehensive chronological overview, see Johnson 2016). Originally planned to monitor the achievement of students aged 9, 13 and 17, and young adults (this intention never materialised), in different subject areas, first surveys were in science, citizenship and writing. Initially, attainment was reported item-by-item, task-by-task, with some items and tasks released for exemplification and others retained securely for re-use in later surveys in order to monitor change over time.

Item-by-item reporting, however, while interesting for teachers and educational researchers, is not particularly useful for policy-makers. Policy-makers need summative information, particularly for evaluating educational initiatives aimed at improving population or subgroup attainment generally. To address policy pressures relating to this and other issues, a managerial move in the early 1980s from the *Education Commission of the States* to the *Educational Testing Service* (ETS) coincided with a programme redesign and the introduction of several procedural changes. Changes included the introduction of a new reporting model, based on subjects rather than on individual items and tasks, an increase in the use of multiple-choice items, and the adoption of IRT for response modelling, analysis and reporting (Messick, Beaton & Lord 1983).

NAEP was originally intended to report attainment at national level only, in response to the concerns of state and local leaders about the possible introduction of a national curriculum, and their fears about likely federal pressure for state-level accountability. However, two decades on from its launch, driven by the report *A Nation at Risk*, NAEP was eventually obliged to begin providing state-level, and even district-level, results in addition to national results, not for explicit accountability purposes but rather to monitor the effectiveness of numerous state-level reforms. Another, inevitable, pressure that NAEP experienced after some years in operation was to change its subject assessment focus to reflect a changing curriculum. Such pressure could have resulted in abandonment of the original NAEP model, with a complete loss of trend data from that point on. Instead, resistance resulted in a decision to run two NAEP programmes in parallel in the future: *Trend NAEP* and *Main NAEP*.

Trend NAEP follows the original NAEP design, and was to be responsible for continuing to document attainment change over time in reading, writing, mathematics and science at the original student ages, using items already used in previous NAEP surveys. In practice, writing assessment was dropped for lack of reliability, and science was eventually also dropped because the original content coverage (the science assessment framework) had become outdated. As a result, trend data over the past four decades is available for reading and mathematics only (National Center for Educational Statistics 2013). In contrast with Trend NAEP, *Main NAEP* is designed to reflect contemporary thinking about what students *should know and be able to do* in a range of subject areas, rather than what they *do know and can do* in a static subset of subjects; the assessment frameworks for Main NAEP are

revised periodically to maintain currency – to embrace interactive digital teaching, for example,. Within Main NAEP there are two component programmes: *National NAEP* and *State NAEP*. The first is based on nationally representative samples of students in Grades 4, 8 and 12, and assesses achievement in a range of subject areas: mathematics and reading are assessed every two years, science and writing are assessed every four years, with other subjects assessed periodically, including the arts, civics, economics, geography, technology and engineering literacy, and US history. The second is based on representative state samples of students in the same grades as *National NAEP*, and assesses achievement in reading, writing, mathematics and science only, in participating states.

Through the 1990s state participation in NAEP was voluntary. It became mandatory for all states following the introduction in 2001 of the *No Child Left Behind Act*. NAEP has in consequence evolved from a programme that furnished national-level attainment information only, of value to federal politicians and policy-makers, to a programme that in addition provides state-level attainment information for use by state as well as federal authorities (see NAEP 2015 for a recent example of 'The Nation's Report Card'). NAEP is not involved, however, in school-level attainment reporting, leaving school-level accountability, with associated performance incentives and penalties (teachers' pay tied to their students' performance) to the states themselves, through their own state-wide exhaustive testing programmes (Hout & Elliott 2011).

4.1.2 Canada

In contrast with the US, system evaluation and monitoring activity in Canada was initially regional only, carried out in individual jurisdictions, in particular British Columbia. It was in the early 1990s that the first nationwide assessment programme in this country, the sample-based *School Achievement Indicators Program* (SAIP), was launched, assessing language, mathematics and science in consecutive years at ages 13 and 16 (Grades 8 and 11), reporting attainment nationally and by jurisdiction. Initially, surveys in science incorporated practical assessment as well as pencil-and-paper assessment, but for financial reasons the practical component was eventually dropped (CMEC 2005). SAIP ran for around a decade, before being replaced by the *Pan-Canadian Assessment Program* (PCAP), which continues in existence today.

PCAP focuses on 13-year-olds only, leaving PISA to assess students later in schooling. The programme is in part modelled on PISA (test construction and administration), and in part on TIMSS (student sampling), to reduce the burden on schools. Like PISA, sample-based surveys take place on a 3-year cycle rather than annually, with language, mathematics and science all assessed on each occasion, one or other subject carrying 'major domain' status each time; every participating student attempts a test booklet containing items from all three domains. Like TIMSS, and unlike PISA, PCAP has adopted a two-stage cluster sampling strategy for student selection: schools are randomly sampled in a first stage and then one class in the target age group is randomly selected within each selected school, with all the students in the selected class selected by default for testing (for the latest survey report, see O'Grady & Houme 2014).

4.2 Europe

Throughout Europe national assessment is either already underway or is planned. As the diversity in purposes, scales and forms is too wide to document fully here, brief case studies for four of the countries with the longest experience of large-scale assessment are offered for illustration.

4.2.1 The UK: England, Wales and Northern Ireland

The UK followed closely behind the US, launching its sample-based APU survey programmes in England, Wales and Northern Ireland in the late 1970s (Foxman, Hutchison

& Bloomfield 1991; Johnson 1989, 2012 Chapter 7, 2016; Newton 2008). In an era characterised by the absence of a national curriculum, the APU focused on language, mathematics and science achievement. In language and mathematics two student age groups were of interest, 11 and 15 years. In science, assessment at age 13 was added; for most students this would be at the end of the second year in secondary school, at which point choices had to be made about which optional subjects, including any sciences, they would begin to study in preparation for external assessment at age 16 in the General Certificate of Education (GCE) O level.

As mentioned in Section 2, the APU programme was replaced in the late 1980s by a cohort-based school accountability model. This initially took the form of a system of NCA based on extensive teacher assessment, to monitor achievement with reference to the newly introduced national curriculum. Excessive teacher workloads led to growing protests in schools, and doubts about the reliability of the national attainment findings among assessment professionals eventually resulted in the first of many changes to the NCA system (Sainsbury 1994), when tests for reading, writing, numeracy and science were introduced. Both the national curriculum and, in particular, the NCA, have continually evolved since (see, for example, Johnson 2012 Chapter 7, 2016). In 2002 the NCA was supplemented by the introduction of a teacher-assessed *Early Years Foundation Stage* (EYFS) profile for 3-5-year- olds.

Currently, England's NCA is based on teacher assessment at KS1 (7-year-olds), and on testing (reading comprehension and mathematics) and teacher assessment (writing) at KS2 (age 11, end of primary school); after being dropped some years ago, science assessment has recently been re-introduced at KS2, with sample-based surveying (Standards Testing Agency [STA] 2017). Assessment at KS3 (13-year-olds), which was essentially test-based, was abandoned entirely a decade ago, following a major disruption in the system when logistic problems associated with the transfer of very large volumes of paper-based scripts from schools to markers to the programme management agency delayed delivery of test results to many schools.

Recent plans to extend the assessment programme, strengthening its school 'value added' data potential by introducing baseline assessment at age 5, foundered when different potential schemes, from different commercial suppliers, were found to produce discrepant attainment results (STA 2016). Government consultation is currently underway with a view to introducing a scheme of baseline teacher assessment for 4-year-olds.

Wales and Northern Ireland were tied to the English model throughout, until the mid-2000s. At this point both countries decided to abandon testing as the basis for national assessment, and to introduce teacher assessment in its place – decisions that brought their own problems in terms of data dependability (Johnson 2013). In light of poor PISA results, and informed also by concerns about the comparability of teacher judgements across the country, Wales has rethought its teacher assessment strategy and reintroduced testing. Northern Ireland, which had used Eleven-plus examination results in place of England's statutory testing, still has no new system in place at KS2, the Eleven-plus system having been discontinued in the meantime.

4.2.2 The UK: Scotland

Scotland, with its historically independent school system within the UK, launched its own sample-based *Assessment of Achievement Programme* (AAP) in the mid-1980s (Condie, Robertson & Napuk 2003; Johnson 2016), modelled to some extent on England's APU. The AAP assessed language, mathematics and science in year groups P4, P7 and S2 (essentially 9-year-olds, 12-year-olds and 14-year-olds, respectively, at the time of testing in May), and reported attainment nationally, initially on an item-by-item, task-by-task, basis.

More than a decade after first introduction, the AAP was successfully remodelled to be in a position to report national student attainment summatively, with reference to the same Level A to F progression framework that teachers had become familiar with after many years of experience with the national 5-14 curriculum (Johnson 1997).

The AAP was 'rebranded' 20 years on, by being replaced by the *Scottish Survey of Achievement* (SSA), which continued in the same vein, but with a broader reporting remit: the SSA was to report attainment summatively by education authority as well as nationally. In addition to language, mathematics and science, a fourth subject, 'social subjects enquiry skills', was added, and the previous 3-year subject cycle changed to a 4-year cycle. The target student year groups were simultaneously changed to P3, P5, P7 and S2, to introduce the possibility of longitudinal age-related progression tracking (e.g., mathematics at P3 to mathematics at P7). In practice, this longitudinal potential, which had existed in the AAP from P4 to P7 (Johnson 1997), was never exploited, partly for reasons of political disinterest, partly for reasons of resource shortage (analysis support), but mainly because the SSA was discontinued before any such additional analyses could be undertaken.

By necessity, the SSA involved an element of cohort testing in small education authorities, which put a high degree of administrative pressure on large schools, as well as inspiring (unsuccessful) demands from the schools inspectorate for school-level performance results, now that these newly existed. For these, and other, reasons, the SSA was replaced in 2011 by the sample-based *Scottish Survey of Literacy and Numeracy* (SSLN).

The SSLN was politically constrained to adopt a stronger practical skills element in its testing, to address the demands of the new skills-focused *Curriculum for Excellence*, despite the likely consequence of a fall in data dependability. Target year groups reverted to P4, P7 and S2 at this point, and a new student sampling strategy was introduced. To facilitate practical assessment in the schools, to increase programme exposure within the teaching profession (a curriculum support motivation), and to avoid future pressure for school-level data from inspectors and others, all schools, rather than a sample of schools, would be expected to participate in surveys, with a mere handful of students tested in each.

Following – some might say giving in to – another international trend, that is the move to school accountability models, the country had at the time of writing just abandoned the SSLN in favour of computer-based 'adaptive' cohort testing at P1 (5-year-olds), P4, P7 and S3. In the event, the new 'standardised testing' programme was abandoned before launch, in face of likely risks to successful implementation (among which were teacher workload issues and information technology readiness in the school system). The country is left with dependence on teacher judgement for system monitoring.

4.2.3 France

France has already been mentioned in Section 2 as another country with a relatively long history of national assessment, though early surveys were not intended for system monitoring. A cohort-testing 'diagnostic' assessment programme was launched in the late 1980s (Bonnet 1997; Trosseille & Rocher 2015), whose purpose was primarily to provide information for school inspectors and receiving teachers about students' strengths and weaknesses as they started a new school year. The French government gathered students' test results from a randomly representative sample of schools for its own analysis purposes, but no analysis results were published. In the early 2000s, however, in response to the 'PISA influence' and growing system accountability expectations, the diagnostic programme was abandoned in favour of an annual cohort testing programme, modelled in part on the NCA in England. There were many problems associated with this programme (see Johnson 2016 for details), and eventually it was abandoned in the early 2010s.

After extensive planning and empirical piloting, a new monitoring and accountability programme – the LOLF (*la loi organique relative aux lois de finances)* – has recently been launched. This is aligned with France's new 'national curriculum' (MEN 2015), and is designed to assess students' achievements in language, mathematics, and elements of science and technology, at the end of the second year of primary school, and at the beginning and end of the lower secondary school (Johnson & Johnson 2016, Section 4.3). Interestingly, the original plan was to assess students at the end of the second and final years of primary school, and at the end of the lower secondary school. A second ambition was to move to online test delivery as the school system became sufficiently well-resourced to accommodate this. As primary schools are not in this position at this time, a relatively late decision was taken to test students at the beginning of the lower secondary school rather than at the end of the primary school. The first survey at the beginning of the lower secondary school took place in November 2015 (for a report see Andreu, Ben Ali & Rocher 2016). A third ambition is eventually to extend the online testing to cover the target populations in their entirety, permitting school-level reporting. Most recently, a change of government has seen the reintroduction of the kind of 'school entry' diagnostic testing that was discontinued in the early 2000s.

Meanwhile, a programme of sample-based subject assessment that began in 2003 continues. Known as CEDRE (*Le cycle des évaluations disciplinaires réalisées sur échantillons*), the programme assesses achievement in French, mathematics, modern languages, civics, science, history, and geography, at the end of primary schooling and at the end of the lower secondary school, with each subject assessed on a six-year cycle.

4.2.4 Hungary

In Section 2 Hungary has been noted as a country with a relatively long history of participation in international survey programmes. Hungary also has many years of domestic programme experience. This began in 1980, with a one-off sample-based survey, *TOF-80*, which assessed achievement in a variety of school subjects in Grades 4 and 8 (Balázsi 2007). This was followed by a sequence of *Monitor Studies*, which ran for around 20 years from implementation in the mid-1980s, at different grade combinations each time. The current programme, the National ABC, was launched in 2001 as a sample-based programme, but in 2008 was transformed into a programme of census surveys of the reading comprehension skills of students in Grades 4, 6 and 8 (Balázsi & Balkányi 2016).

4.3 Australasia

4.3.1 New Zealand

As noted in Section 2, New Zealand was among the handful of countries with the earliest national assessment programmes. In the case of New Zealand this was the sample-based NEMP, which was launched in the mid-1990s, focused on year groups 4 (8-9 year-olds) and 8 (12-13 year-olds), and ran for 15 years (Crooks & Flockton 1993; Flockton 2012). NEMP was notable for its inclusion of a broad range of school subjects in its survey programme, for its use of 'rich assessment tasks' in the subjects assessed, and for its determined focus on reporting performance at the level of individual items and tasks only, for the benefit of the teaching profession rather than policy-makers. Among the areas of the curriculum surveyed by NEMP were reading, writing, mathematics, science, technology, social studies, physical education, and health, art, and music. Different subject groups were assessed each year, and re-assessed every three or four years. Depending on the subject domain, the assessment materials used included pencil and paper tests, interviews, videos, performance-based tasks, small science experiments, dramatization with puppets, producing art works, singing and dancing, and physical agility (Flockton 2012).

NEMP was discontinued when the government's 15-year commissioning contract with the University of Otago ended. This coincided with the point at which National Standards were

introduced, and a new cohort-based programme involving teacher assessment of students' reading, writing and mathematics achievement began to be introduced.

4.3.2 Australia

A relative newcomer to national assessment, Australia currently has two programmes in operation: the cohort-based *National Assessment Program – Literacy and Numeracy* (**NAPLAN**) and *National Assessment Program Sample Assessments* (NAP-SL, science literacy; NAP-CC, civics and citizenship; NAP-ICTL, ICT literacy).

NAPLAN began in 2008, and is a programme of annual national cohort assessment for students in Years 3, 5, 7 and 9, with testing in May. Four areas of development are assessed each time – reading, writing, language conventions (spelling, grammar, punctuation) and numeracy. Prior to 2016, NAPLAN testing was referenced to national 'Statements of Learning' for English and for Mathematics. These were developed collaboratively in the early 2000s by State, Territory and Australian education authorities, to address concerns about the lack of curriculum consistency that then existed in the various jurisdictions across the country, by defining and delivering common curriculum outcomes to inform curriculum development in those jurisdictions. Over time the 'Statements' were absorbed into Australia's first national curriculum, and in 2016 NAPLAN testing was aligned with the newly revised standards-based Australian National Curricula for English and for Mathematics. Phased over a 2-3 year period, NAPLAN is scheduled to move from its current paper-based delivery model to online delivery from 2017, with 'tailored testing' for increased efficiency.

The NAP sample assessments run on a 3-year cycle, monitoring students' skills and understanding in science literacy, civics and citizenship, and information and communication technology (ICT) literacy in Year 6 and, with the exception of science literacy, in Year 10. Science literacy is assessed in Year 6 only, leaving PISA to provide relevant attainment information for Year 10. The programme began in 2003 with a survey of science literacy, followed in 2004 with civics and citizenship, and in 2005 with ICT literacy, with that pattern repeating thereafter. Surveys in civics and citizenship have been delivered online since 2013, in ICT literacy since 2014, and in science literacy since 2015.

Schools have access to their own students' attainment results, and parents have access to their own children's performance results.

4.4 Africa

Many countries in eastern and southern Africa have participated in one or more of the cross-national SACMEQ surveys, as outlined in Section 2.2. Some have also set up their own national assessment programmes in parallel, not all of which have adopted SACMEQ's sample-based approach. South Africa is a particularly interesting example.

4.4.1 South Africa

South Africa launched its programme of *Annual National Assessments* (ANA) in 2011, and ended it in 2016. The ANA was a cohort testing programme that focused on literacy and numeracy, reporting attainment at national, provincial, district and school levels for each of Grades 1 to 9 (RSA 2015). Testing was carried out in several different languages, with annual reports produced for each skill area, in each language, at each grade. The programme initially covered Grades 1 to 6, with a planned progressive grade inclusion of Grade 9 in 2012 and Grades 7 and 8 in 2015, bringing the annual testing load to over 8.5 million leaners. Perhaps not surprisingly, given the scale and relatively rapid implementation of the programme, there were problems.

Many issues arose, including cost, dependability of results, and teacher workload, with teachers threatening and implementing boycotts as the number of grades tested rose and the number of individual students tested increased. Despite a generally positive stakeholder consensus that the ANA served a useful purpose in highlighting strengths and weaknesses in achievement in the schools, negative reactions to the experience in the field led to a government review of ANA practice (purpose, frequency, scope, standard and quality of tests, reliability of outcomes, and utilisation of results by schools), a public consultation on a new draft policy for the ANA, and postponement of the 2015 survey into early 2016.

In May 2017 the then Basic Education Minister announced that the ANA was to be replaced by the *National Integrated Assessment Framework* (NIAF), with associated diagnostic tests introduced for the benefit of class teachers, complemented by summative examinations and 'independent systemic evaluations'. The latter will focus on Grades 3, 6 and 9, will be sample-based, and conducted on a 3-year cycle starting in 2018.

4.4.2 Uganda, Malawi, and Zimbabwe

Uganda, Malawi and Zimbabwe, like South Africa, all gained large-scale assessment experience through their participation in SACMEQ cross-national surveys, and have consolidated that experience through domestic initiatives. Uganda, for example, has for some time been benefitting from its own domestic programme, the *National Assessment of Progress in Education* (NAPE). This operates in both the primary and secondary sectors (Grades 3 and 6, and Senior Grade 2), with annual sample-based surveys of literacy and numeracy in the former, and English, mathematics and biology in the latter (World Bank 2012a). With USAID support, Malawi carried out EGRA surveys at Standards 2 and 4 in 2010, 2011 and 2012, the last of these based on large nationally representative samples of students, with each student individually orally assessed (Pouezevara, Costello & Banda 2013). In addition, with financial support from UNICEF, the country launched its cohort-based *Monitoring Learning Achievement* (MLA) programme in 2012, at Grades 2, 4 and 7; this is planned to run on a 3-year cycle, assessing mathematics, Chichewa language and English language each time. For its part, Zimbabwe ran a four-year programme, the *Zimbabwe Early Learning Assessment* (ZELA), one of whose aims was to evaluate the effectiveness of UNICEF's *Education Development Fund* activity in the country; the programme focused on the language and numeracy skills of children beginning Grade 3 in primary school (UNICEF-ACER 2016).

4.4.3 Rwanda

Rwanda has not yet participated in any regional initiatives involving large-scale assessment. Nevertheless, without the capacity building benefit of such participation, but with UNESCO support (REB 2012), the country launched its own domestic programme in 2011, the sample-based *Learning Achievement in Rwandan Schools* (LARS). LARS focuses on literacy and numeracy in the primary school, using paper-based assessment materials. While the first survey, LARS 1, tested children in P3, with reference to achievement in the national curriculum, the second and third switched attention to P2 and P5. Among other problems, there have been issues to do with sample sizes, and inadequate numbers of test items in tests to support over-time monitoring.

In the meantime, as part of a move to harmonise education systems across its partner states of Burundi, Kenya, Rwanda, South Sudan, Tanzania, and Uganda, whilst bringing them into line with international trends, the *East African Community* (**EAC**) has recently finalised a 'competence-based' *2013 Harmonised Curriculum Framework for the East African Community.* Rwanda has in response remodelled LARS to align with the new EAC curriculum. This has entailed changing the age groups once again, to align with those planned for a possible future regional large-scale assessment programme.

4.4.4 Ethiopia, Ghana, and Sudan

Several other African countries that are not members of regional consortia have nevertheless gained experience in large-scale assessment. Ethiopia, Ghana, and Sudan are just three examples. With donor support, and every 3-4 years since the end of the 1990s, Ethiopia has conducted its *Ethiopian Baseline National Learning Assessment* (EBNLA) at Grades 4 and 8 in a variety of subjects; in 2009 the country carried out the *Ethiopian First National Learning Assessment* (EFLNA) at Grades 10 and 12 (World Bank 2009). With USAID funding support, Ghana has been running its sample-based NEA biannually since 2005, assessing English language and mathematics in Grades 3 and 6 (World Bank 2013a). For its part, Sudan has recently (2009-11) carried out sample-based surveys in some states at Grades 4 and 5, and was set to scale up to national surveys of reading in Grade 3 by 2016, by launching its *National Learning Assessment* (NLA) programme (World Bank 2013b), with funding support from the Global Partnership for Education within its Basic Education Recovery Project.

4.5 Latin America

Since the early 1990s, in addition to participating in LLECE surveys, several countries in the region have launched their own independent domestic survey programmes, with mixed success. At least two countries – Guatemala and Venezuela – actually discontinued their newly launched programmes when international donor support ended (Ferrer & Fiszbein 2015). In other countries, national assessment programmes remain in operation, albeit with evolutions to respond to changing policy demands.

4.5.1 Chile

Chile is noted in Section 2 as one of the handful of countries in the world that have the longest histories of large-scale assessment experience, dating back to the 1980s (Johnson 1999; Ferrer 2006; Gysling 2016). National assessment in this country has not been particularly benign, however. As Gysling (2016: 20) notes, 'assessment in Chile has been used historically as a policy tool by the state to implement its educative programme, and not just in the present'. In the 1980s Chile launched its *Programa de Evaluación del Rendimiento* (PER), assessing students in Grades 4 and 8 in language, mathematics, natural sciences and social sciences. The PER eventually became the *Sistema de Medición de la Calidad de la Educación* (SIMCE), which continues today (Meckes & Carrasco 2010; Gysling 2016). SIMCE is a system of cohort assessment, with a high-stakes school accountability role: for example, SIMCE school attainment results are used to allocate competitive funds for educational improvement projects and serve as indicators for school incentivisation (Ferrer 2006).

One of the changes made when the direction of the education agenda was modified in light of previous national assessment results, and the SMCE introduced, was to:

> *… align the national curriculum and assessment so as to ensure coherence in educational policy and to encourage the implementation of the new curriculum in classrooms.* (Gysling 2016: 18)

4.5.2 Brazil

Brazil has a more recent history of national assessment. With resource support from the UNDP, Brazil piloted its sample-based *Sistema de Avaliação da Educação Básica* (SAEB) in the early 1990s, to assess the achievement of students in Grades 5 (end of primary school), 9 (end of lower secondary school) and 12 (end of secondary school) in reading (Portuguese) and mathematics; occasionally other subjects were also assessed, including history, geography, and science (Guimarães de Castro 2012). After a pause for reflection and modification, and with World Bank funding, the programme was launched nationally in 1995 and ran for almost a decade on a 2-year cycle (Canen 2012; Guimarães de Castro 2012; Paget, Malmberg & Martelli 2016). In 2005, the programme was renamed *Prova Brasil*, and

in urban areas cohort testing was introduced, opening the way for school-based accountability; sample-based assessment continued in rural areas and in the private sector (Canen 2012). Eventually, many states and municipalities also put in place their own large-scale assessment programmes.

4.6 The Middle East

4.6.1 Jordan

Jordan embarked on a programme of education reform in the early 1990s, but its most comprehensive 10-year reform programme – *Education Reform for Knowledge Economy* (ERfKE) – was launched in 2003, with multi-donor support. Improving the ways that learning outcomes are measured was one of the main interventions of the teaching and learning component of ERfKE. After gaining experience from participation in international surveys, the country developed and implemented its own national assessment programmes to monitor reform impact (Obeidat & Dawani 2014, Chapter 4). A programme of national cohort testing, the *National Test for the Control of Education Quality* (NRCEQ) was launched in 2000 and focused on three grades each year (4, 8 and 10). The programme was reformed in 2004 to align with the newly reformed curriculum, and from this point on the focus shifted to a cyclic pattern, with a single grade tested each year rather than three. To better detect and monitor achievement trends over the course of the education reform process, a second national assessment programme was established. Sample-based surveys within the *National Assessment for Knowledge Economy* (NAfKE) were carried out in 2006, 2008 and 2011.

With USAID support, Jordan, along with Egypt and a number of African countries (Gambia, Kenya, Liberia, Malawi, Mali, Mozambique, and South Africa), has also recently introduced EGRA into the early grades of primary school (Jordan simultaneously introduced EGMA), and launched national surveys to gauge impact. The first EGRA and EGMA surveys were carried out in Jordan in 2012, with second surveys planned for 2014 (Obeidat & Dawani 2014). Egypt's first EGRA was implemented in 2013 (LaTowsky, Cummiskey & Collins 2013), and was followed with a second survey in 2014 (RTI International 2014).

4.6.2 Qatar

Qatar launched its *Qatar Comprehensive Educational Assessment* (QCEA) in 2004, with the collaboration of the RAND Corporation, which the Government of Qatar had invited to evaluate its current education system and to recommend change. The QCEA involves cohort testing at every grade in Qatar's primary and secondary school systems (Brewer et al. 2007), and has proved problematic for this and other reasons (Gonzalez et al. 2009; RAND 2009). Since the school year 2005-06, Qatar has been publishing an annual compendium of statistics relating to its school system, including students' attainment results in language, mathematics and science in Grades 4 to 11 - the most recent report offers information for the 2014-15 school year (MEHE 2016).

4.6.3 Bahrain and UAE

With support from Cambridge International Examinations, Bahrain began a programme of cohort-based national assessment in 2009, with programme roll out planned eventually to cover Grades 3, 6, 9 and 12 (Cambridge International Examinations 2015). For its part, the UAE launched its national assessment programme, the UAENAP, in 2010, with support from the Australian Council for Research in Education (ACER) (Egbert 2012). The programme assessed Arabic, English, mathematics and science in Grades 3, 5, 7 and 9, with cohort testing in the Emirates of Ajman, Dubai, Fujairah, Ras Al Khaimeh, Sharjah, and Umm Al Quwain. One policy consequence of the survey findings was a planned overhaul of the curriculum generally.

4.6.4 Saudi Arabia

Saudi Arabia was slow to embark on the road to national assessment, but the country has just experienced what must be one of the fastest implementations of surveys in a planned long-term national assessment programme. Senior figures in the country were only recently lamenting the continuing unavailability of such a programme for monitoring and improving the quality of the, still very traditional, Saudi education system (e.g., Al Sadaawi 2010). In response, the country's *Public Education Evaluation Commission* (PEEC) developed a 7-year strategic plan for national assessment, and in 2015, in cooperation with ACER, implemented sample-based surveys in Grades 3 and 6 (middle and end of primary school), to be followed in 2016 with surveys at these same stages and in Grades 4 and 5 (ACER 2016b). The programme is planned for further expansion each year, in terms of grades and subjects surveyed.

4.7 Central and South East Asia

4.7.1 Vietnam, Laos and Cambodia

In 2000, with support from the World Bank, UNESCO and the UK's Department for International Development (DFID), among other donors and aid agencies, the Vietnam Ministry of Education and Training (MoET) launched a large-scale monitoring study of primary education, the *Reading and Mathematics Assessment Study*. Surveys were conducted in 2001 (World Bank 2004) and 2007 in Grade 5, and in 2009 at Grades 6 and 9. In Laos the *National Assessment of Student Learning Outcomes* (ASLO) saw its first survey (ASLO I) conducted in 2006, in public schools in the primary sector, with a repeat in 2009 (ASLO II) which covered private schools as well. ASLO III followed in 2012, two grades lower in the primary sector, while ASLO IV is planned for 2017 (RIES 2015). More recently, as mentioned earlier, in 2011-12 Vietnam, Laos, and Cambodia participated in large-scale assessment in the primary sector, within the French-language PASEC programme (for findings from the Vietnam surveys and Cambodia's 'diagnostic evaluation' in four primary grades, see, respectively, PASEC 2014a, 2014b).

4.7.2 Kazakhstan

Kazakhstan launched its first national assessment programme in 2005. This was the Interim State Control (ISC), which ran for six years, and which comprised annual census surveys of student achievement in Grades 4 and 9 across a rotating set of school subjects. In 2011, in response to the adoption of a *State Program for Education Development*, the ISC was replaced with the sample-based *External Assessment of Learning Achievement* (EALA), whose first survey focused on Grade 9 and assessed the language of instruction, the history of Kazakhstan, algebra and chemistry (World Bank 2012b).

4.7.3 Afghanistan

One of the latest countries in this region to embark on introducing a national assessment system is Afghanistan, which, with ACER support, launched its *Monitoring Standards in Educational Growth* (MTEG) programme in 2013, with a sample-based survey of Class 6 students (last year in primary school) in government schools in 13 Afghan provinces taught in Dari or Pashto; survey findings are reported in Lumley et al. (2015). The MTEG is designed as a long-term monitoring programme with a focus on trends in achievement outcomes in key stages over time (Classes 3, 6 and 9), and another focus on learning progression from Class 3 through Class 6 to Class 9. When fully in place, the MTEG will assess each key stage in schooling on a 3-year cycle. "It is envisaged that the program will expand to implementation in other countries" (Lumley et al. 2015: 3).

4.7.4 SEAMEO regional initiative

Plans are advanced for the launch of a possible future cross-border survey programme involving the 11 member countries of **SEAMEO** (*Southeast Asian Ministers of Education Organization)*: Brunei Darussalam, Cambodia, Indonesia, Laos, Malaysia, Myanmar,

Philippines, Singapore, Thailand, Timor-Leste, and Vietnam. With UNICEF and ACER support, a contextualised *South East Asia Primary Learning Metric* (SEA-PLM) has been developed for use at Grades 4-5 (10-year-olds) in SEAMEO countries, with a focus on literacy, numeracy and global citizenship (ACER 2016a). Field trials took place in a small subset of countries during 2015-16, with a first full survey across all the SEAMEO countries planned for 2017. A similar future development for use with 7-year-olds is apparently a possibility.

5. In conclusion: issues, trends, and dilemmas

5.1 Political, economic, logistic, technical, impact and other issues

The previous section will have provided an indication of the extent of national assessment activity around the world, and a flavour of the variety of models that are in operation across different countries, large and small. Occasionally, issues raised during programme planning and implementation have been noted.

The difficulties faced in practice when one-off surveys or longer term monitoring programmes are implemented are in principle many, but problems are rarely recorded in survey reports, programme evaluations, or research articles. That said, a few accounts of problematic practice do exist that identify some issues that could be resolved and others that proved insurmountable, and sometimes fatal, in terms of original programme goals and design. The information offered here draws on accounts from a number of such sources: Ferrer 2006, in the context of national assessment in Latin America; Green, Bell, Oates and Bramley 2008, reviewing issues surrounding national curriculum assessment in England; Gonzales et al. 2009, and RAND 2009, documenting programme development experience in Qatar; Kellaghan, Bethell and Ross 2011, offering general guidance on issues in large-scale assessment; Kuan 2011, reflecting on experience in Egypt and other developing countries; Flockton 2012, on experience in New Zealand; Guimarães de Castro 2012, on experience in Brazil; Bakker 2014, evaluating large-scale computerised adaptive testing in Georgia; Obeidat and Dawani 2014, offering lessons learned from Jordan; Tobin et al. 2015, based on experience in the Asia-Pacific region; Johnson 2016, analysing developed world case studies from the perspective of 'intelligent accountability'; and UNICEF-ACER 2016, providing an account of experience in Zimbabwe.

Issues can be loosely grouped into several broad categories, including but not necessarily limited to the following:

- Political control, commitment and interference
- Infrastructural inadequacies
- Resource shortfalls
- Technical challenges
- School overload
- System impact
- Programme management.

5.1.1 Political control, commitment, and interference

Among the various stakeholder groups, national politicians have the greatest power over national assessment, as explained in Section 1. It is they who decide whether to newly launch, change the nature of, or to render defunct a national assessment programme. It is they also who generally decide the purposes of the enterprise, and hence the goals to be achieved, the resources to be made available, the management structure to be adopted, and so on. Politicians tend to have a short-term view in most things, however, when a system monitoring programme should by definition be a tool for the long term. They have been known to replace or to drastically modify an existing well-functioning monitoring programme when a new government is elected, naively expecting then to see rapid improvements in 'attainment standards' during their personal terms of office. These issues of political control, questionable commitment and programme interference are arguably among the most serious problems in this field.

5.1.2 Infrastructural inadequacies

Infrastructural inadequacies arise principally in developing countries, where access to schools can be difficult, and postal and telecommunication services might be unreliable or non-existent. A problem in both developing and developed countries is lack of access to a comprehensive and reliable programme-relevant educational management information system (EMIS), with which to carry out school and student sampling, and to record events as each survey and the programme as a whole progresses. Where programme planners have been ready to move from paper-based testing to an internet-delivery model the computer infrastructure in schools has frequently been found wanting.

5.1.3 Resource shortfalls

The quality of the assessment materials that are administered to students within any assessment exercise is crucial in assuring the validity, interpretability and utility of the resulting attainment findings. Quality materials need to be available in time and in sufficient quantity for use in a first survey, to represent adequately the target domain and its subdomains. Further, they need to be available at this time in sufficient volume to furnish materials for use in subsequent surveys as well (see Section 5.3 for further discussion on this point), including the calibrated 'link items' required in IRT applications. In countries where students might attempt tasks and tests in different languages within a survey, establishing materials equivalence in the different languages is an additional challenge.

Inadequate programme resourcing is a recurring issue in the field, which goes beyond the availability of suitable high-quality assessment materials. Financial resourcing is a particular problem in developing countries, and lack of economic resource has proved fatal in some countries when international donor support has ended. In developed countries, available budgets, along with logistic problems, have generally precluded the inclusion of the assessment of practical skills in surveys, jeopardising the validity of the reported findings in terms of curriculum impact and system effectiveness. Skills shortfalls must be mentioned alongside financial issues. Even where budgets could cover the cost of an appropriate level of human resource to manage and implement assessment programmes successfully, some of the skills required are not readily available. A criticism made by some observers is that even where the right skills are available when a programme is launched, those skills are often lost when the programme is abandoned, leaving a skills vacuum when a new programme is eventually launched.

5.1.4 Technical challenges

The principal technical challenges that are faced in programme design concern sampling, and data analysis and reporting. Where a programme is sample-based it is essential that the sampling strategy adopted produces an intended sample of students that faithfully represents, or intentionally misrepresents (with deliberate under-or over-sampling of student subgroups), the target student population. It is equally essential that when the survey is implemented the nature of the achieved sample in important respects is recorded, so that appropriate data weighting can be applied to redress unwanted imbalances, and also to aid interpretation of any observed attainment change.

Sampling is also an issue as regards item development, whether pretesting is designed to furnish empirical information about item performance for general screening purposes (including validity checking) or for later use in test creation and analysis (item calibration in IRT models). One can add to this the technical know-how that is necessarily required for data analysis and reporting, particularly in IRT applications. Solving technical issues principally requires access to the appropriate high-level technical skills of statisticians and psychometricians, especially where programmes are sample-based and/or use IRT for attainment analysis and reporting.

5.1.5 School overload

Cohort testing systems impose great pressure on schools, particularly when class teachers not only organise and supervise test sessions, but also mark students' scripts. In systems where the cohort testing is carried out annually in several consecutive school stages, the negative impact on schools and teachers can quickly become overwhelming, and has led to wholesale protest and even test boycotts in some of the countries that practice this. The pressures are heavier in paper-based assessment contexts. Another pressure on some schools occurs when they are selected to participate in an international survey, and this is carried out in the same year as a national survey, perhaps in the same stage and at closely similar times of year.

5.1.6 System impact

All stakeholder groups can be expected to be impacted in some way, by the results that emerge from national assessment programmes. There is evidence from some countries, both developed and developing, that this is indeed the case, as far as policy-makers, school inspectors, and managers are concerned. Among teachers the picture is patchy. Those teachers that have been involved in implementing attainment surveys, whether as item writers, test administrators, or markers, have benefitted from their involvement. Reaching the teaching profession more widely, however, continues to be a problem – addressing this by involving every teacher actively in some way in surveys is not a feasible proposition, and becomes less so with the increasing use of internet test delivery and automated marking.

As far as the impact of programme findings on education systems is concerned, again it is difficult to assess whether lessons have been learned and put into practice or not. System monitoring programmes probably do have an impact on national policy-making, but this is not always recorded for public consumption. And there is policy being made that is claimed to be informed by national, and international, survey results, and that in practice is simply being justified by them in retrospect. Where programmes might be used to evaluate the effect of policy initiatives on student attainment, the short lifespan of many programmes has meant that this potentially invaluable evaluation tool is unavailable for the purpose.

5.1.7 Programme management

The need for distributed programme management, with responsibilities shared among organisational partners, and the equally important need for adequate resourcing, both financial and human, are lessons that should, and could, be addressed in future programmes, both in developed and developing countries – in the latter by country governments and international sponsors jointly. These problems recur.

5.2 Trends in focus, scale and methodology

Even as national assessment activity has been rapidly increasing worldwide over the past two decades, a number of trends in practice are readily observable. Principal among these are the focus of assessment, the scale of assessment, and the methodology of choice.

5.2.1 Assessment focus

When national and international assessment programmes first emerged, attention focused on specific 'key' curriculum subjects, including language and mathematics as universals, but also subjects such as science and geography. One of the fastest-growing trends in recent years has been a move to the assessment of 'cross-curricular skills', principally literacy and numeracy, or, more correctly, reading comprehension and numeracy, given that writing is now rarely assessed. An associated trend is the almost exclusive adoption of objective items in large-scale assessment. PISA's recent excursion into 'problem-solving' is a notable exception that is still under trial.

There are several drivers that explain these rapid and widespread trends. One is the increasing impact of national accountability agendas, stimulated and supported by the OECD, and underwritten by the shared assumption that a literate and numerate population leads to strong national economic growth. A second influential factor is the general embrace of technology in schools, with its motivational influence on students (in learning and in assessment), and its potential for lower cost, logistically less challenging, student assessment: through, for example,, elimination of the high costs associated with the preparation and transportation of paper-based materials, and of the high workload implications of human marking when automated marking is available.

5.2.3 Scale of assessment

Arguably, the most important aspect of survey design, one that results in important differences in a programme's ability to meet some specific intended purposes, is that of a sample-based approach versus cohort assessment. Sample-based surveys and census surveys can both meet most of the purposes listed in Section 1 to some more or less acceptable degree. There is just one potential purpose for national assessment that a sample-based approach cannot satisfy to the extent that cohort testing can, and this is the question of school accountability.

School-level accountability is such a serious matter for schools and their staff, whether or not rewards and sanctions are applied for 'good' or 'poor' performance, that between-school comparisons cannot be left to sample estimates, even where within-school student samples are representative and relatively large. This, along with the scarcity of sampling expertise, explains the current rapidly increasing move from sample-based surveys to cohort assessment, and in particular to cohort assessment in sequential, and even consecutive, school stages (intended in principle to explore age-related and school-related progression). But cohort assessment is not necessarily the answer to policy demands for accountability data, as explained in Section 5.3.

5.2.4 Measurement methodology

A third strong trend, that is supported by continually improving school computerisation, and the ready availability of appropriate software and training, from the IEA and other agencies, is the move to the adoption of IRT methodologies in national assessment. Online adaptive testing is a particularly sophisticated example. This move is rendered feasible not only by school computerisation and internet access, but also by the concomitant moves to a narrow focus on reading and numeracy assessment, as mentioned above, and associated use of mainly objective items that lend themselves to automated marking.

IRT methodologies, though, are not always entirely appropriate for use in this context. This is not simply because of their sophistication, and sometimes inscrutable 'black box' data manipulations, but more fundamentally because of an important assumption underpinning the validity of outcome interpretation. The Rasch model in particular, can be problematic for this reason. The underpinning assumption of this model is that of 'item invariance'. This means that, while any two test items can change their level of difficulty over time, in response to national teaching influences perhaps, or can be different from one student subgroup to another at any point in time, their *relative* difficulties remain the same for every student group within and over time (or approximately so, within arbitrary limits). Where items exhibit 'differential item functioning' in pretesting, they are generally removed for survey use, potentially threatening the validity of the set that remains, in terms of curriculum or skill set representation.

5.3 The accountability dilemma

The recent and growing popularity of cohort testing systems, many replacing long-established sample-based programmes, has arguably exacerbated the problems associated with multiple, non-prioritised and sometimes conflicting programme goals. Further, when combined with annual testing in several consecutive school grades, cohort testing has impacted negatively on the workload of teachers and damaged their goodwill and programme buy-in. Replacing a 'one test for all' approach with matrix sampling, and eliminating the need for class teachers to supervise test sessions and to mark students' scripts, would reduce the pressure on schools. But this would be at increased cost, if not computerised, and would probably be excluded on that basis. For those countries with a strong and dependable computer infrastructure, the internet delivery of tests and automation of item marking could be a solution, but this strategy, too, will offer inevitable problems, not least in terms of reducing assessment validity in the short term. Combined with issues of cost and logistics, cohort assessment clearly precludes any real possibility of addressing performance skills within surveys, compromising the validity of subject assessment where practical skills feature strongly.

Cohort testing is assumed to facilitate school accountability. Yet the validity of between-school performance comparisons is not beyond question in this context. This is for two main reasons. Firstly, in order for school comparisons to be seen to be fair, all the students in every school across the country will probably be required to attempt the same test(s) in the assessed subject domain(s) in any year, unless very sophisticated and relatively untried IRT strategies, including adaptive testing, are used. Since tests are usually kept short to avoid overburdening students, and disrupting normal school activity, curriculum or cross-curricular skills representation within the test will not be comprehensive. There will therefore be assessment validity issues here at the level of individual students. Also, where entire classes of students are assessed simultaneously, sitting in close proximity to one another in conditions that might be far from the controlled conditions of high-stakes examinations, cheating, intentional or otherwise, could be too difficult to resist, reducing further the likely validity of the individual outcomes.

Moreover, where a school population contains very small schools, such as primary schools in rural areas, the student cohort can be quite changeable from one year to another, so that school performance in one year could in principle be quite different the following year. Even in large secondary schools sudden, or even almost imperceptibly gradual, changes in school catchment characteristics can cause interpretational difficulties. Hence the need for relatively sophisticated 'value added' approaches to inter-school comparisons. Having the test in several different versions, modified by item presentation order, or using different 'equivalent' tests administered with matrix sampling, could solve this problem, but both strategies would be more expensive and difficult logistically to implement.

So cohort testing is not necessarily the perfect approach that politicians and policy-makers assume it to be, for providing interpretable school-level attainment data within and over time. The risks of non- valid interpretations of between-school attainment differences and progression are not insignificant.

A dilemma for programme designers and others, and one which might not yet be recognised as such, is how much effort and cost they are prepared to invest in cohort assessment, as opposed to sample-based assessment? How important is the need for school-level attainment data? Does the imperative for school accountability data override the possibility of providing better cross-sectional sample-based attainment data for monitoring population attainment over time?

References

[All available web links – offered behind author name(s) and publication date – were accessed in October 2017]

ACER (2014). *The Latin-American Laboratory for Assessment of the Quality of Education: Measuring and comparing educational quality in Latin America*. Assessment Gems Series, no.3. Melbourne: Australian Council for Research in Education.

ACER (2016a). *Southeast Asia Primary Learning Metrics. Audit of Curricula*. Melbourne: Australian Council for Research in Education.

ACER (2016b). A collaborative approach to national assessment in Saudi Arabia. *International Developments*, 6, article 6.

Al Sadaawi, A.S. (2010). Saudi National Assessment of Educational Progress (SNAEP*). International Journal of Education Policy and Leadership*, 5(11).

Andreu, S., Ben Ali, L. & Rocher, T. (2016). *Évaluation numérique des compétences du socle en début de sixième : des niveaux de performance contrastés selon les académies.* [Computer-based assessment of national curriculum skills at the beginning of secondary school: Comparative attainment levels by authority.] Note d'Information, no.18. Paris: Ministry of Education.

Baird, J-A., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T. & Daugherty, R. (2011). *Policy Effects of PISA*. Oxford University Centre for Educational Assessment.

Baird, J-A., Johnson, S., Hopfenbeck, T., Isaacs, T., Sprague, T., Stobart, G. & Yu, G. (2016). On the supranational spell of PISA in policy. Educational Research, 58(2): 121-138.

Bakker, S. (1999). Educational Assessment in the Russian Federation. *Assessment in Education: Principles, Policy & Practice*, 6(2): 291-303.

Bakker, S. (2014). *The Introduction of Large-scale Computer Adaptive testing in Georgia*. A READ Publication.

Balázsi, I. (2007)**.** *Results of the National Assessment of Basic Competencies in Hungary*. Background paper prepared for the *Education for All Global Monitoring Report 2008 Education for All by 2015: will we make it?*

Balázsi, I. & Balkányi, P. (2016). *Comparing Results of PIRLS and the Hungarian National Assessment of Basic Competencies.* Paper presented at the *European Conference on Educational Research* (ECER), Dublin, 23-26 August.

Bialecki, I., Johnson, S. & Thorpe, G. (2002). Preparing for National Monitoring in Poland. *Assessment in Education: Principles, Policy & Practice*, 9: 221-236.

Bohla, H.S. (2012). Social and Cultural Contexts of Educational Evaluation: A Global Perspective. In T. Kellaghan, D.L. Stufflebeam & L. Wingate. (eds), *International Handbook of Educational Evaluation. Part One: Perspectives*, pp397-416. Springer.

Bonnet, G. (1997). Country profile from France. *Assessment in Education: Principles, Policy and Practice*, 4, 295-306.

Bradshaw, J., Sturman, L., Vappula, H., Ager, R. & Wheater, R. (2007). *Achievement of 15-year-olds in Wales: PISA 2006 National Report.* Slough: National Foundation for Educational Research.

Brewer, D.J., Augustine, C.H., Zellman, G.L., Ryan, G., Goldman, C.A., Stasz, C. & Constant, L. (2007). *Education for a new era: design and implementation of K–12 education reform in Qatar. Executive Summary*. RAND Corporation.

Canen, A. (2012). Assessment of schools in Brazil: some reflections. *SA-eDUC JOURNAL,* 9: 1, 1-8.

Chinapah, V. (1995). *Monitoring Learning Achievement. Towards Capacity Building*. Final Report. Paris: UNESCO.

Cambridge International Examinations (2015). *Working with ministries of education and national examination boards.* Cambridge: Cambridge International Examinations.

CMEC (2005). *School Achievement Indicators Program. SAIP, Science 2004*. Toronto: Council of Ministers of Education, Canada.

Condie, R., Robertson, I. J. & Napuk, A. (2003). The Assessment of Achievement Programme. In T. G. K. Bryce, & W. M. Humes (eds), *Scottish Education.* Edinburgh: Edinburgh University Press.

Crooks, T. J. & Flockton, L. C. (1993). *The design and implementation of national monitoring of educational outcomes in New Zealand primary schools.* Dunedin, New Zealand: Higher Education Development Centre.

Egbert, A. (2012). A clearer picture: national and international testing in the UAE. *International Developments*, 2.

Eurydice (2009). *National testing of pupils in Europe: Objectives, organisation and use of results.* Brussels: Education, Audiovisual and Culture Executive Agency.

Ferrer, G. (2006). *Educational Assessment Systems in Latin America: Current Practice and Future Challenges.* Washington, DC: PREAL.

Ferrer, G. & Fiszbein, A. (2015). *What has happened with learning assessment systems in Latin America? Lessons from the Last Decade of Experience.* Commission for Quality Education for All Background Paper. Washington, DC: The Dialogue.

Flockton, L. (2012). *The Development of the Student Assessment System in New Zealand.* Washington, DC: The World Bank.

Foxman, D., Hutchison, D. & Bloomfield, B. (1991). *The APU Experience 1977-1990.* London: Schools Examination and Assessment Council.

Gonzalez, G., Le, V-N., Broer, M. Mariano, L.T., Froemel, J.E., Goldman, C.A. & DaVanzo, J. (2009). *Lessons from the Field. Developing and Implementing the Qatar Student Assessment System, 2002-2006.* The Rand Corporation.

Greaney, V. & Kellaghan, T. (eds) (2008). *Assessing National Achievement Levels in Education*. Washington, DC: The World Bank.

Green, S., Bell, J.F., Oates, T. & Bramley, T. (2008). *Alternative Approaches to National Assessment at KS1, KS2, and KS3*. Cambridge: Cambridge Assessment.

Guimarães de Castro, M.H. (2012). *Developing the Enabling Context for Student Assessment in Brazil.* Washington DC: The World Bank.

Gysling, J. (2016). The historical development of educational assessment in Chile: 1810–2014, *Assessment in Education: Principles, Policy & Practice*, 23(1): 8-25.

Hout, M. & Elliott, S.W. (eds) (2011). *Incentives and test-based accountability in education.* Washington, DC: The National Academies Press.

Husén, T. (ed.) (1967*). International Study of Achievement in Mathematics: a comparison of twelve countries*. Stockholm: Almqvist & Wiksell.

Husén, T. & Postlethwaite, T. N. (1996). A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education: Principles, Policy & Practice,* 3, 129-142.

Jeantheau, J-P. & Johnson, S. (2016). *Literacy in France. Country Report*. *Short Version*. European Literacy Policy Network (ELINET).

Johnson, S. (1989). *National Assessment: The APU Science Approach*. A technical report on programme development. London: HMSO.

Johnson, S. (1997). Issues in National Assessment: the AAP. In Kirkwood, M.J., Roger, A. & Rideout, P. (eds), *Proceedings of the 1996 SERA Conference*. Glasgow: University of Strathclyde.

Johnson, S. (1999). International Association for the Evaluation of Educational Achievement Science Assessment in Developing Countries. *Assessment in Education: Principles, Policy & Practice*, 6(1): 57-73.

Johnson, S. (2012). *Assessing learning in the primary classroom*. London: Routledge.

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education*, 28:1, 91-105.

Johnson, S. (2016). National assessment and intelligent accountability. In Wyse, D., Hayward, L. & Pandya, J. (eds), *SAGE Handbook on Curriculum, Pedagogy and Assessment*. Chapter 53, Part 5: Assessment and the Curriculum. London: Sage Publications.

Johnson, S. & Johnson, R. (2016). *Literacy in France. Country report. Children and adolescents.* European Literacy Policy Network (ELINET).

Jones, L. V. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher*, 25, 1-8.

Kellaghan, T., Bethell, G. & Ross, J. (2011). *Assignment Report: Guidance Note: National and International Assessments of Student Achievement*. HDRC, UK.

LaTowsky, R.J., Cummiskey, C. & Collins, P. (2013). *Egypt Grade 3 Early Grade Reading Assessment Baseline*. United States Agency for International Development.

Lumley, T., Mendelovits, J., Stanyon, R., Turner, R. & Walker, M. (2015). *Class 6 proficiency in Afghanistan 2013 : outcomes of a learning assessment of mathematical, reading and writing literacy.* Melbourne: Australian Council for Educational Research.

Makuwa, D. (2010). *What are the levels and trends in reading and mathematics achievement?* SACMEQ Policy Issues Series, Number 2. Paris : SACMEQ/IIEP.

Martin, M. O., Mullis, I. V. S., Foy, P. & Hooper, M. (2016). *TIMSS 2015 International Results in Science.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Meckes, L. & Carrasco, R. (2010). Two decades of SIMCE: An overview of the national assessment system in Chile. *Assessment in Education: Principles, Policy & Practice,* 17, 233–248.

MEHE (2016). *Education in the Schools of the State of Qatar. Annual Report for the Academic Year 2014/2015.* State of Qatar Ministry of Education and Higher Education.

MEN (2015). *Le socle commun de connaissances, de compétences et de culture*. Bulletin Officiel, no.17, avril. Paris: French Ministry of Education.

Messick, S., Beaton, A. & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era*. Princeton: National Assessment of Educational Progress.

Mirazchiyski, P. (2013*).* ***Providing School-Level Reports from International Large-Scale Assessments: Methodological Considerations, Limitations, and Possible Solutions*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics.* Boston College, TIMSS & PIRLS International Study Center.

Mullis, I.V.S., Martin, M.O., Foy, P. & Drucker, K.T. (2012). *PIRLS 2011 International Results in Reading.* Chestnut Hill, MA, USA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

National Center for Educational Statistics. (2013). *The nation's report card: Trends in academic progress 2012 (NCES 2013 456)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

NAEP (2015). *The Nation's Report Card. Science 2015*.

Newton, P. E. (2008). *Monitoring national attainment standards.* London: Office for Qualifications and Examinations Regulation.

Obeidat, O. & Dawani, Z. (2014). *Disseminating and Using Student Assessment Information in Jordan*. Washington, DC: The World Bank.

OECD (2001). *Knowledge and Skills for Life: First Results from PISA 2000*. Executive summary. Paris: OECD Publications.

OECD (2009). *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*. Paris: OECD Publishing.

OECD (2014). *PISA 2012 Results in Focus: what 15-year-olds know and what they can do with what they know*. Paris: OECD Publications.

OECD (2016a). *PISA 2015 Results in Focus*. Paris: OECD Publications.

OECD (2016b). *PISA 2015 Results. Excellence and Equity in Education. Volume I*. Paris: OECD Publications.

O'Grady, K. & Houme, K. (2014). *PCAP 2013. Report on the Pan-Canadian Assessment of Science, Reading, and Mathematics.* Toronto: Council of Ministers of Education, Canada.

Paget, C.L., Malmberg, L-E. & Martelli, D.R. (2016). Brazilian national assessment data and educational policy: an empirical illustration. *Assessment in Education: Principles, Policy & Practice*, 23:1, 98-125.

PASEC (2014a). *School Performance and Factors of Public Primary Education Quality in the Kingdom of Cambodia. Diagnostic Evaluation Report - Cambodia 2011/2012.* Dakar: CONFEMEN.

PASEC (2014b). *School Performance and Factors of Public Primary Education in the Socialist Republic of Vietnam*. Dakar: CONFEMEN.

PASEC (2015). *PASEC2014. Education System Performance in Francophone Sub-Saharan Africa. Competencies and learning factors in primary education*. Dakar: CONFEMEN.

Peaker, G. F. (1975). An empirical study of education in twenty-one countries: a technical report. New York: Wiley.

Pellegrino, J. W. (2014). *National Assessment of Educational Progress*.

Pouezevara, S., Costello, M. & Banda, O. (2013). *Malawi National Early Grade Reading Assessment Survey. Final Assessment – November 2012.* United States Agency for International Development.

Purves, A.C. (1991). Brief history of IEA. In W. A. Hayes (Ed.), *Activities, institutions and people. IEA Guidebook 1991.* The Hague: The International Association for the Evaluation of Educational Achievement.

RAND (2009). *Lessons Learned from Developing and Implementing the Qatar Student Assessment System*. Research Brief. Rand-Qatar Policy Institute.

REB (2012). *Learning Achievement in Rwandan Schools (LARS).* Kigali: Rwanda Education Board, Education Quality and Standards Department.

Rey, O. (2010). The use of external assessments and the impact on educational systems. In S.M. Stoney (Ed.), *Beyond Lisbon 2010: perspectives from research and development for education policy in Europe*. (CIDREE Yearbook 2010). Slough: NFER.

RIES (2015). *Lao People's Democratic Republic - National Assessment of Student Learning Outcomes.* Laos Government Research Institute for Educational Science.

RSA (2015). *Action Plan to 2019. Towards the Realisation of Schooling 2030.* Pretoria: Republic of South Africa.

RTI International (2014). *Egypt Grade 3 Early Grade Reading 2nd National Assessment*. United States Agency for International Development.

Sainsbury, M. (1994). The structure of National Curriculum Assessment. In Hutchison, D. & Schagen, I. (eds), *How Reliable is National Curriculum Assessment?* Slough: National Foundation for Educational Research.

Sargent, C., Foot, E., Houghton, E. & O'Donnell, S. (2013). *INCA Comparative Tables. International Review of Curriculum and Assessment Frameworks Internet Archive (INCA)*. London: Department for Education.

Scheerens, J., Ehren, M., Sleegers, P. & de Leeuw, R. (2012). *OECD review on evaluation and assessment frameworks for improving school outcomes. Country background report for the Netherlands*.

Shiel, G., Kavanagh, L. & Millar, D. (2014). *The 2014 national assessments of English reading and Mathematics. Volume 1: Performance Report*. Dublin: Educational Research Centre.

Silova, I. & Khamsi, G. (2008). Introduction: Unwrapping the Post-Socialist Education Reform Package. In Silova, I. & Steiner-Khamsi, G. (eds), *How NGOs React: Globalization and Education Reform in the Caucasus, Central Asia and Mongolia.* Bloomfield, CT: Kumarian Press.

SKBF/CSRE (2014). *Swiss Education Report 2014*. Aarau: Swiss Coordination Centre for Research in Education.

Spaull, N. (2011). *Primary School Performance in Botswana, Mozambique, Namibia, and South Africa.* Working Paper 8. SACMEQ.

Spencer, E. (2013). National Assessments: Improving Learning and Teaching through National Monitoring? In T.G.K. Bryce, T.G.K., W.M. Humes, W.M., Gillies, D. & Kennedy, A. (eds), *Education in Scotland*. Edinburgh: Edinburgh University Press.

Statistics Commission (2005). *School Education Statistics: User Perspectives*. Report no. 26. London: Statistics Commission.

STA (2016). *Reception baseline comparability study. Results of the 2015 study*. London: Standards and Testing Agency.

STA (2017). *Key stage 2 science sampling 2016. Methodology note and outcomes.* London: Standards and Testing Agency.

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R. & Nyamkhuu, T. (2015). *Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific region.* Melbourne: ACER and Bangkok: UNESCO.

Trosseille, B. & Rocher, T. (2015). Les évaluations standardisées des élèves. Perspective historique. *Education & Formations*, 86-87: 15-36.

UNESCO (1990). *World Declaration on Education For All. Framework for Action to meet Basic Learning Needs*. Paris: UNESCO.

UNESCO (2000a). *World Education Forum. Final Report*. Paris: UNESCO.

UNESCO (2000b). *The Dakar Framework for Action. Education for All: Meeting our Collective Commitments*. Paris: UNESCO.

UNESCO (2015). *Incheon declaration. Education 2030: Towards Inclusive and equitable quality education and lifelong learning for all*. Paris: UNESCO.

UNICEF-ACER (2016). *Zimbabwe Country Case Study, 2016*.

West, R. & Crighton, J. (1999). Examination reform in Central and Eastern Europe: issues and trends. *Assessment in Education: Principles, Policy & Practice*, 6(2): 271-290.

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

Wilson, J. W. & Peaker, G. F. (eds) (1971). International study of achievement in mathematics. *Journal for Research in Mathematics Education*, 2, special issue.

World Bank (2004). *Vietnam Reading and Mathematics Assessment Study*. Volume 3. Washington, DC: The World Bank.

World Bank (2009). *Ethiopia. Student Assessment*. SABER country report.

World Bank (2012a)**.** *Uganda. Student Assessment*. SABER country report.

World Bank (2012b). *Kazakhstan. Student Assessment*. SABER country report.

World Bank (2013a). *Ghana. Student Assessment.* SABER country report.

World Bank (2013b). *Sudan. Student Assessment.* SABER country report.

Wyse, D. & Torrance, H. (2009). The development and consequences of national curriculum assessment for primary education in England*. Educational Research*, 51(2): 213-228.