ACTRC

Large-Scale Assessments for use in the Philippines

Assessment Curriculum and Technology Research Centre
University of the Philippines, Diliman, Quezon City 1101, Philippines
www.actrc.org

Australian Aid

THE UNIVERSITY OF MELBOURNE

UNIVERSITY OF THE PHILIPPINES 1908

The Assessment, Curriculum And Technology Research Centre is a partnership between The University of Melbourne and The University Of The Philippines supported by the Australian Government.

List of Acronyms

ASER - Annual Status of Education Report

EALAS - East Asia Learning Achievement Study

EGMA - Early Grade Mathematics Assessment

EGRA - Early Grade Reading Assessment

LAMP – Literacy Assessment and Monitoring Programme

LAPG - Language assessment for Primary Grades

LLECE - Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (Latin American Laboratory for Assessment of the Quality of Education)

LMTF - Learning Metrics Task Force

NAT - National achievement test

NCAE - National Career Assessment Exam

PASEC - Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (Programme on the Analysis of Educational Systems)

PIRI - The Philippine Informal Reading Inventory

PIRLS - Progress in International Reading Literacy Study

Pre-PIRLS - Pre-Progress in International Reading Literacy Study

PISA - Programme for International Student Assessment

READ - Russian Education Aid for Development

SACMEQ - Southern and Eastern Africa Consortium for Monitoring Educational Quality

SEA-PLM - South East Asia Primary Learning Metric

SEAMEO - Southeast Asian Ministers of Education Organisation

TIMSS - Trends in International Mathematics and Science Study

UNESCO - United Nations Educational, Scientific and Cultural Organisation

UNICEF - United Nations Children's Fund

Table of Contents

List of Tables and Figures

Identity and Acknowledgements

This report has been prepared on behalf of the Assessment Curriculum and Technology Research Centre, in response to a request from the National Assessment Technical Working Group on System Assessment. The contracting authority is the University of Melbourne Commercial. The technical work has been completed by staff of the Assessment Research Centre of the Melbourne Graduate School of Education, University of Melbourne.

The team is grateful for the input of its affiliated staff at the Assessment Curriculum and Technology Research Centre, UP Diliman; from members of the Department of Education, Philippines; and from professionals in the Philippines who have previously been involved in the country's large-scale assessment initiatives.

Project Lead	Esther Care PhD
Project Team	Farhan Azim
	Bruce Beswick PhD
	Susan-Marie Harding PhD
	Rebekah Luo PhD
Review Team	Therese Bustos PhD
	Louie Cagasan Jr

1 Introduction

1.1 Basis of review

Stakeholders in education development, including Ministers of Education and policy makers, are increasingly looking to data regarding student learning outcomes in order to inform evidence-based policy decisions. Measuring learning outcomes at the national and international level can provide policy makers with information with which to diagnose the strengths and weaknesses of educational programs, and inform educational reform. Analysis of the data can explain the factors that may contribute to student growth within and across countries.

The number of countries participating in national, regional and international assessments has grown rapidly (Benavot & Tanner, 2007; Ritzen, 2013). Associated with this growth has been an increase of research into the impact of national and international assessment programs on education policy, teaching and learning practices (Best et al., 2013).

Wagner et al. (2012) claim that the usefulness of an assessment is contested "particularly … in low-income countries where the growth in assessments is most rapidly expanding and where the empirical base is least developed" (p. 510). The authors indicate that the usefulness of an assessment for a particular country depends on:

a) *who* gets tested
b) *what* gets tested
c) *when* tests occur
d) *how* a test takes place
e) *why* a test takes place.

This review has as its focus assessments that can be used for international comparisons. Therefore, assessment programs that are specific to one research initiative, or assessment tools available from commercial companies, are not included. While these types of materials are useful for teachers and schools and the data produced can inform the student, parent, teachers and schools on student achievement, there is no reference point to student achievement in other countries. Similarly, international assessments that no longer operate are not included in this review.

There are many large-scale assessment initiatives that provide information about student achievement. These include international assessment programs (open to all countries), regional or multi-country programs (restricted to a limited number of countries or a geographic area), national programs, household surveys, hybrids (involving both large-scale assessment tools and household survey tools), and system strengthening programs. These assessment types, with examples, are shown in Figure 1.1.

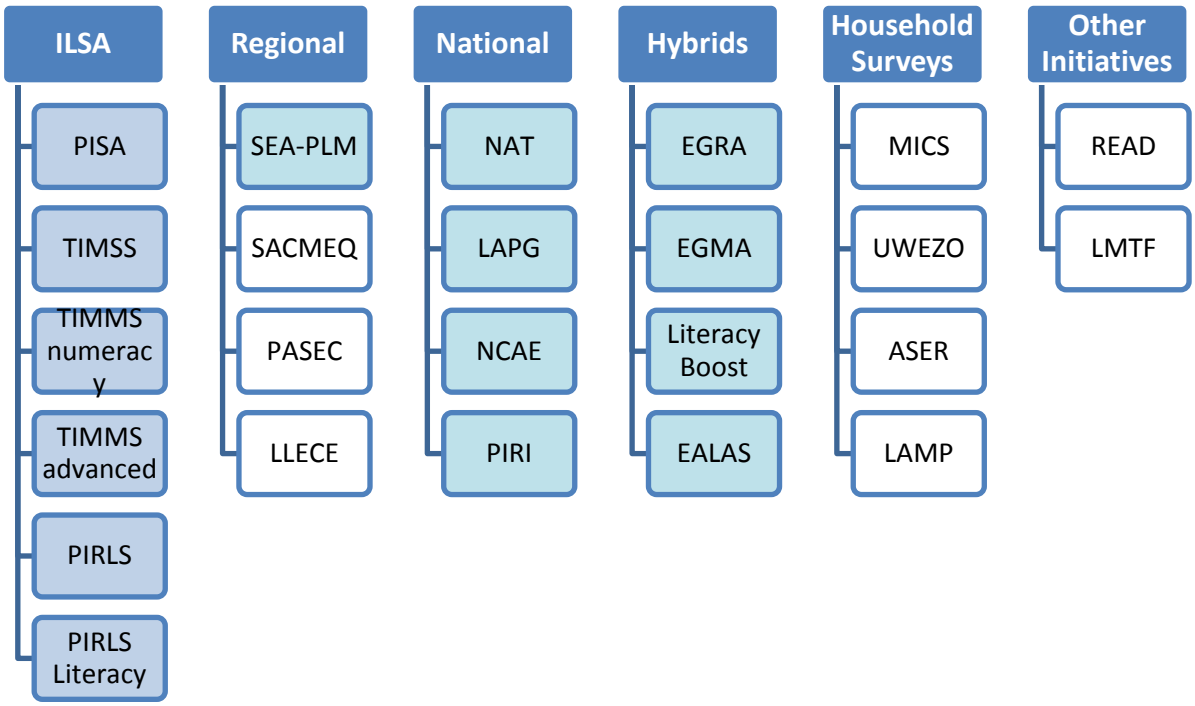

Figure 1.1 Types of Assessment Programs and Initiatives

Note. Blue shade – possible Philippine involvement; pale blue shade – Philippine involvement; white shade – not possible involvement.

1.2 Philippines education reform

In 2011, the Philippine Department of Education initiated a curriculum reform which resulted in the implementation of the Enhanced Basic Education (K to 12) Program. The aim of this program is to raise the Philippines' elementary and secondary educational standards. The program covers mandatory kindergarten for 5-year-old children, followed by six years of primary education, four years of junior high school and two years of senior high school. The rationale for this reform is to provide sufficient time to master expected competencies outlined in the national curriculum, develop lifelong learners, and prepare graduates for tertiary education, middle-level skills development, employment and entrepreneurship. The new curriculum concentrates on six salient features (Government of the Philippines, 2014):

1. Strengthening Early Childhood Education (Universal Kindergarten)
2. Building Proficiency through Language (Mother Tongue-Based Multilingual Education)
3. Making the Curriculum Relevant to Learners (Contextualisation and Enhancement)
4. Ensuring Integrated and Seamless Learning (Spiral Progression)
5. Gearing Up for the Future (Senior High School)
6. Nurturing the Holistically Developed Filipino (College and Livelihood Readiness, 21st Century Skills).

The roll-out of the new curriculum will be completed by the end of the 2017-2018 school year (SY). Students who enter Grade 1 in SY 2012-2013 will complete the full K to 12 curriculum upon high school graduation in 2025. At the end of SY 2017-2018, the Grade 12 graduates will be the first to complete the enhanced secondary education program (Figure 1.2).

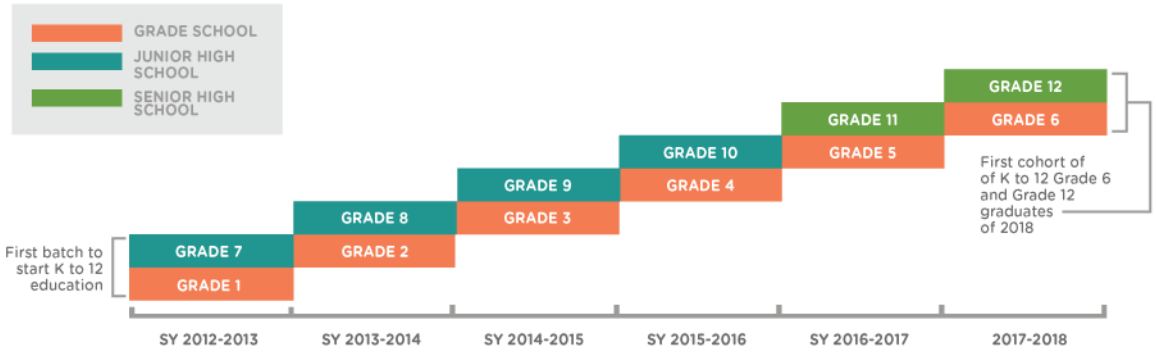

Figure 1.2 Dates of roll-out of K–12 Curriculum (www.gov.ph/K-12)

The Congress of the Philippines, Republic Act No. 10533 (Republic of the Philippines, July 2012), states the "state shall create a functional basic education system that will develop productive and responsible citizens equipped with the essential competencies, skills and values for both life-long learning and employment. In order to achieve this, the state shall:

a) Give every student an opportunity to receive quality education that is globally competitive based on a pedagogically sound curriculum that is at par with international standards
b) Broaden the goals of high school education for college preparation, vocational and technical career opportunities as well as creative arts, sports and entrepreneurial employment in a rapidly changing and increasingly globalized environment, and
c) Make education learner-orientated and responsive to the needs, cognitive and cultural capacity, the circumstances and diversity of learners, schools and communities through the appropriate languages of teaching and learning, including mother tongue as a learning resource."

The imperative that the state will provide an education that is globally competitive immediately places the onus on the state to provide evidence to inform progress toward that goal. One form of evidence can be obtained through benchmarking of Philippine student progress against other countries, using international and/or multi-country assessments. A major issue which might confound the interpretation of such benchmarking data concerns the time taken for the K to 12 curriculum implementation to filter down to consequentially increased student outcomes. There is no set time period for a new curriculum to be considered adequately implemented such that effects on student outcomes are measurable (Care & Beswick, 2016).

The education community uses large-scale assessment data to evaluate progress in learning and education. Data are used in different ways according to national priorities and governance. This review is undertaken with the assumption that the purpose of use of large-scale assessment data in

the Philippines would be to provide the country with evidence of progress toward its goals as outlined in a) to c) above. The focus on progress (as opposed only to meeting of goals) is important to note, and is consistent with the point made that lag-time between implementation of a reform and its achievement is unknown.

Note that the use of assessment data as a baseline measure before a country implements system-wide change has not been thoroughly reviewed by any of the international large-scale assessment authorities. Using national assessment data as a baseline measure has been reviewed by the World Bank (Greaney & Kellanghan, 2008). The degree to which the Philippines can rely on its current and previous national assessments to evaluate progress is minimal, due to the fact that the implementation of the K to 12 reform is well underway, and comparability of previous assessment data with current is tenuous given that both the curriculum and the assessments that are aligned with these differ. Comparability relies on common curriculum, common assessment, and/or common students.

1.3 Use of large-scale assessments

In this review both international and regional assessments are considered. The Philippine national assessments are referred to for timing purposes and administration considerations, but these assessments have not been reviewed.

For the purposes of this review 'large-scale assessment' is defined as measurement of student learning designed to describe the achievement of students in particular areas of learning across an education system. The term 'international large-scale assessment' is used to describe such an assessment across different countries not defined by a region. The term 'regional large-scale assessment' refers to assessment across different countries within a particular geographic region, and is a subset of international large-scale assessment programs. Large-scale assessment can be implemented through population or sample approaches. A 'population approach' refers to assessment of all students within the target range in a participating country; a 'sample approach' refers to assessment of selected students within the target range. In the latter condition, an assessment can be administered to a (selected) sample of students and findings can be extrapolated statistically to describe the population.

International large-scale assessments (ILSA) provide data on several countries; thereby countries' education systems' outcomes can be compared using students' results. An issue for ILSA is the degree to which assessment tools function similarly across countries due to language, culture and education system differences. To assess student achievement within a country in terms of that country's specific educational goals, national assessment tools will be the most appropriate. Where, however, ILSA targets the same educational goals, results from these will be of major interest. The decision by a country to participate in ILSA needs to be made in the light of the latter's capacity to provide information of interest to that country.

There is mixed information concerning the degree to which countries use ILSA outcomes to inform policy and reform; and the degree to which the technical capacity building that occurs consequential upon participation in ILSA, is sustained and useful.

1.3.1 Informing policy

Tables 1.1 and 1.2 show the performances and rankings in TIMSS and PIRLS of a number of countries and administrative regions that have used ILSA results to inform education policy with the aim of improving student achievement. As the tables show, Hong Kong has been a very strong performer in TIMSS mathematics and a relatively strong performer in TIMSS science since 1995. In PIRLS reading literacy, however, its 2001 result was regarded in the region as disappointing, particularly because it followed the implementation in 1994 of a program to improve language education. This relatively disappointing result became the subject of discussion in Hong Kong's government and a number of new strategies were subsequently adopted to improve language education outcomes. Talks and workshops were organised to disseminate the PIRLS results to schools and to instruct parents on the establishment of a good home reading environment. PIRLS reading skills began to be included in the region's curriculum, and in 2004 the PIRLS framework was adopted across the region for the Chinese reading comprehension examination. The improvement in the next PIRLS assessments was remarkable: the region's average score improved from 528 in 2001 (17th) to 564 in 2006 (2nd) and 571 in 2011 (1st).

Table 1.1 TIMSS rankings for a selection of countries using ILSA data to inform policy

Assessment Year	**1995**		**1999**	**2003**		**2007**		**2011**		**2015**	
Education Systems	**4th grade**	**8th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**
Hong Kong											
Mathematics	**587**	**588**	**582**	**575**	**586**	**607**	**572**	**602**	**586**	●	●
Rank	4th	4th	4th	2nd	3rd	1st	4th	3rd	4th		
Science	**533**	**522**	**530**	**542**	**556**	**554**	**530**	**535**	**535**	●	●
Rank	10th	16th	15th	4th	4th	3rd	9th	9th	8th		
New Zealand											
Mathematics	**499**	**508**	**491**	**493**	**494**	**492**	-	**486**	**488**	●	●
Rank	13th	15th	21st	17th	20th	23rd	-	31st	16th		
Science	**531**	**525**	**510**	**520**	**520**	**504**	-	**497**	**512**	●	●
Rank	11th	15th	19th	12th	13th	22nd	-	31st	15th		
Norway											
Mathematics	**502**	**503**	-	**451**	**461**	**473**	**469**	**495**	**475**	●	●
Rank	12th	17th	-	21st	27th	25th	21st	29th	20th		
Science	**530**	**527**	-	**466**	**494**	**477**	**487**	**494**	**494**	●	●
Rank	12th	14th	-	20th	21st	25th	18th	33rd	19th		
Russian Federation											
Mathematics	-	**535**	**526**	**532**	**508**	**544**	**512**	**542**	**539**	●	●
Rank	-	11th	12th	9th	12th	6th	8th	10th	6th		
Science	-	**538**	**529**	**526**	**514**	**546**	**530**	**552**	**542**	●	●
Rank	-	9th	16th	9th	17th	5th	10th	5th	7th		
South Africa											
Mathematics	-	**354**[1]	**275**	-	**264**	-	-	-	**352**[2]	-	●
Rank	-	41st	38th	-	45th	-	-	-	44th	-	
Science	-	**326**[1]	**243**	-	**244**	-	-	-	**332**[3]	-	●
Rank	-	41st	38th	-	45th	-	-	-	45th	-	
Total	39	41	38	27	47	37	50	50	45	48	60

● = participation but data not yet available

- = no participation or no data available

[1] = used unapproved sampling procedures and did not meet other guidelines
[2] = 9th grade tested and average achievement not reliably measured because the percentage of students with achievement too low for estimation exceeded 25%
[3] = reservations about reliability of average achievement because the percentage of students with achievement too low for estimation does not exceed 25% but exceeds 15%
Results drawn from Reports on International Achievement in Mathematics and Science, 1995-2011, http://timssandpirls.bc.edu/

Table 1.2 PIRLS rankings for a selection of countries using ILSA data to inform policy

Assessment Year	2001	2006	2011	2016
Education Systems	**4th grade**	**4th grade**	**4th grade**	**4th grade**
Hong Kong				
Reading	**528**	**564**	**571**	●
Rank	17th	2nd	1st	
New Zealand				
Reading	**529**	**532**	**531**	●
Rank	13th	24th	23rd	
Norway				
Reading	**499**	**498**	**507**	●
Rank	25th	35th	31st	
Russian Federation				
Reading	**528**	**565**	**568**	●
	16th	1st	2nd	
South Africa				
Reading	-	**302**	-	●
Rank	-	45th	-	
Total	35	45	45	52

● = participation but data not yet available
- = no participation or no data available
Results drawn from Reports on International Achievement in Reading, 2001-2011, http://timssandpirls.bc.edu/

The case of Hong Kong usefully illustrates both the advantages and dangers of using large-scale assessments to inform education policy. Tjeerd Plomp, former chairman of the governing body of TIMSS and PIRLS – the IEA – has said that the organisation initially "held back on reporting raw scores, because such scores easily make a study into a kind of 'Olympics' or 'horse race'" (Plomp, 1992, p. 282). Rankings and other simple comparisons attract media headlines but can obscure differences of cultural significance in national education curricula (Care & Beswick, 2016). The governing bodies of large-scale assessments deliberately focus on testing only skills and knowledge that are shared by participating countries (e.g. http://www.oecd.org/pisa/aboutpisa/pisafaq.htm). Nevertheless, those bodies recognise and even promote the usefulness of ILSA data for the reform of education policy. As Plomp (1992, p. 279) has said, "the IEA collects the sort of data policy-makers can use as a basis for decision-making to improve education."

Despite some negative issues, unfavourable media headlines can play a role in education policy development. As the TIMSS and PIRLS International Study Centre (2011) has noted, evidence of underperformance often spurs educational reform. One danger, however, is that a rush to respond to news of disappointing results may lead to the short-term solution of 'teaching to the test.' While sample items are an acceptable part of test preparation, a deeper understanding of the general principles that underlie the items allows for the application of the required skills to problems outside

the testing environment and creates a stronger foundation for future learning. Hong Kong's response to its 2001 reading literacy result combined the practising of PIRLS questions with other initiatives. Notably, the improvement was achieved despite low levels of interest in reading among students and their parents. Even in 2011, when the region ranked 1st out of 45 countries in reading literacy, its students were ranked 39th in their liking for reading, 42nd in their level of engagement by reading lessons, 44th in their reading confidence, and 45th in their motivation to read. Even more remarkable in the context of Hong Kong's campaign to create better home reading environments, the region's parents ranked 45th in their interest in reading (Tse et al., 2012). These results may say something about the demands of the region's new reading education program, but they also indicate that strong progress can be made when the importance of an educational achievement is recognised and focused initiatives are implemented.

Another striking example of improved PIRLS performance, and one that almost exactly parallels the Hong Kong example, is that of the Russian Federation. As Table 1.2 shows, the country's average score improved from 528 in 2001 (16th) to 565 in 2006 (1st) and 568 in 2011 (2nd). Particularly notable were its students' improved performances in the interpretation of texts, the integration of ideas and information, and the analysis and evaluation of content and language. Froumin and Kuznetsova (2012) conclude that these improved results should not be attributed to any one factor but to a combination of factors. These include an increase in the average age of primary school students, an increase in the proportion of entry-level students considered by parents and principals to be 'school-ready,' and an increase in the average socioeconomic status of the students' families. Another likely contributor was a structural change in the primary school system that took place between the two assessments. In 2001, 63% of primary school students attended schools that taught Grades 1 to 3 only, and 37% attended schools that taught Grades 1 to 4 only. In 2006, almost all the students who participated in PIRLS were at schools that taught Grades 1 to 4. It is possible that the greater continuity created by this structural change played a significant role in bringing about the improved results of 2006. However, when the improvement in analytical, interpretive and evaluative skills is taken into account, it seems likely that a qualitative reform in the nature of Russian education also played an important role. From the mid-1990s, the focus of Russian education began to shift from 'reproductive' teaching methods aimed at imparting skills and knowledge in a 'ready-made' form to more active and creative methods aimed at facilitating students' ability to direct their own learning. Froumin and Kuznetsova (2012) suggest that this reform had become embedded by the time of the 2006 PIRLS assessment, resulting in a student cohort less focused on content regurgitation and more focused on the critical and creative skills required for developmental learning.

However, not all reform is good reform, and the need to identify effective strategies is underlined by mixed results in other countries that have used ILSA data to inform their education policies. Among these countries the TIMSS and PIRLS International Study Centre identifies Norway, which has developed education policy in response to various international assessments since 2000. The country uses ILSA data to evaluate the skills and knowledge of its students in relation to those of other countries, to develop benchmarks and to set national policy. Its strategies have included a focus on basic skills in reading, mathematics and science, increased education for teachers, and a National Quality Assessment System. Following the 2006 PIRLS results, Norway established a program in early commencement for reading instruction, early interventions for weak learners, and a renewed emphasis on reading throughout primary education. As Tables 1 and 2 show, however,

the country's TIMSS and PIRLS results indicate no clear pattern of improvement. Between 1995 and 2011 its achievements in mathematics and science declined and then fluctuated but made no overall gains. In the PIRLS assessment, the country's achievement in reading literacy remained steady between 2001 and 2006, though it declined in relation to other countries (from 25th to 35th). In 2011 it improved marginally, while in relation to other countries it occupied a position between its two previous rankings. These results can in part be attributed to fluctuations in the numbers of countries participating in the various assessments but, overall, no substantial improvement was recorded.

Another country that has purposely developed education policies in response to ILSA data is South Africa. Following the 1999 and 2003 TIMSS assessments, in which the Republic's average scores were 243 (38th of 38 countries) and 244 (45th of 47 countries), it initiated reforms throughout its education system, directing resources towards mathematics and science, and deciding that performance on TIMSS would serve as a measure of the effectiveness of its reforms. Its aim was to use TIMSS data to identify areas of weakness in its education system and to use the results as a benchmark to measure school effectiveness. The country did not participate in the 2007 TIMSS assessment but when it returned to TIMSS in 2011 its average score in mathematics was recorded as 352 (44th of 45 countries) and its average score in science was recorded as 332 (45th of 45 countries). These figures appear to show improvement from 1999 and 2003 but, as the footnotes to Table 1.1 indicate, the 2007 mathematics test was administered to the country's 9th grade rather than its 8th grade, and there is doubt about the reliability of both results due to the relatively high percentage of participating students with achievement too low for estimation. Whether the country's program of resources and reforms had an overall positive effect on the performance of its students is uncertain. This result may indicate that progress in ILSAs depends on the ways in which resources and reforms are managed, or it may indicate that reforms and resources can be ineffective when broader social and economic challenges, such as high concentrations of poverty and crime, are too great. Other countries with social and economic challenges, such as the territories governed by the Palestinian Authority, have also fared poorly in ILSA results (Reports on International Achievement in Mathematics and Science, 1995-2011, http://timssandpirls.bc.edu/).

While higher socioeconomic status is correlated with higher educational outcomes, research has identified a law of diminishing returns in educational resourcing (Betts, 1999). Beyond a certain level that has been achieved in most developed countries, increases in resourcing have little impact on student performance. An interesting education system to observe in the immediate years ahead will be that of New Zealand, a country with high socioeconomic status and a government that uses ILSA data for system-level monitoring and evidence-based policy development. New Zealand has reviewed achievement as measured by ILSAs to determine alignment with the national standards applied in its primary schools, and its Ministry of Education (2011) has released a Statement of Intent for 2011/12–2016/17 identifying the goals of improved literacy and numeracy in comparison to other countries as measured by PISA, TIMSS and PIRLS. It has also sought to improve educational outcomes through its Iterative Best Evidence Synthesis Program, a collaborative knowledge-building strategy designed to strengthen the evidence base that informs education policy and practice in the country. As Table 1.1 shows, the average scores of New Zealand's 4th grade students in mathematics and science declined gradually from 1995 to 2011 while their performances in comparison to other countries declined more dramatically, from 13th in 1995 to 31st in 2011. A similar pattern is evident in the country's reading performances, as shown in Table 1.2. The average score actually improved

marginally from 2001 to 2006 despite declining dramatically in relation to other countries, falling from 13th in 2001 to 23rd in 2006, before stabilising in 2011. Performances of the country's 8th grade students in mathematics and science fluctuated between 1995 and 2011, both in their averages and in relation to other countries, but made no overall improvement. These data illustrate the dangers of emphasising comparisons to other countries rather than comparisons over time within a country, but their relative stability as average scores provides a useful baseline from which to begin the measurement of change under a reform program. With a stable economy and minimal social unrest in the country, New Zealand's achievements in ISLAs in the years ahead may provide useful evidence for the efficacy of its reforms.

The assumption that improvement in large scale assessment programs is evidence of actual educational standards improvement needs to be explored. Specific questions to consider include:

- Does improvement of student performance as evidenced in ILSA, in areas such as literacy and numeracy, imply that actual standards in these target areas, as evidenced through national and school level goals, have also improved?
- Does improvement of student performance as evidenced in ILSA, in areas such as literacy and numeracy, imply that standards in other subjects that might draw on these skills have also improved?
- Does improvement of student performance as evidenced in ILSA, in areas such as literacy and numeracy, imply that standards in education more generally have also improved?

1.3.2 Technical capacity building

As will be apparent from the following descriptions of ILSA programs, they require different levels of input from participating countries in terms of technical contributions. These contributions typically include:

- Fieldwork during pilot and program phases
- Provision of data about the education provision in the country in terms of numbers of schools and students, as well as about factors of interest for use in sampling
- Implementation of the assessments including management of security, distribution of assessments and collection of assessments
- Within country training of assessment administration personnel
- Within country training of personnel to mark open-ended questions.

In some ILSA, in addition, more technical contributions are required, for example:

- Drawing of the sampling frame
- Decisions concerning replacement of schools within the sampling frame
- Analysis of country data and reporting.

Each of these activities imply the acquisition and use of in-depth technical expertise around test and scale development, assessment, and data analysis. The teams of personnel that are required for implementation may receive both training through international workshops, within country professional development and training, and the accompanying experience. This training is a valuable resource upon which to draw particularly in developing countries where the skills pool in this area is

scant. The learnings can be applied by in-country personnel to their national and regional assessment initiatives.

The training is valuable at country-level only when it is sustained, consolidated, and leads to flow-on capacity building to contribute to the country's assessment expertise. ILSA provides the context and trigger for this, but the sustainability relies on the conscious decision of the country to build upon the experience through incorporating the learning and the personnel into the country's organisational units that are dedicated to both system and regional assessment provision and maintenance.

2 Profiles of international large-scale assessments

2.1 Programme for International Student Assessment (PISA)

2.1.1 Summary and aims

The Organisation for Economic Co-operation and Development (OECD) runs a triennial international survey, the Programme for International Student Assessment (PISA). The goal of PISA is to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students, who in a majority of systems are approaching the end of their compulsory education. PISA tests how students can apply their knowledge to real life situations and problems, rather than testing their knowledge recall. "PISA looks at students' ability to apply knowledge and skills in key subject areas and to analyse, reason and communicate effectively as they examine, interpret and solve problems" (PISA, 2015a).

The Directorate for Education at the OECD manages PISA, while representatives from participating countries are involved through the PISA Governing Board. The Directorate contracts out the technical aspects related to the assessments to institutions[1] with the technical expertise to manage these. PISA results are analysed on a national level, extrapolating to show countries where they stand against other countries. To date there have been five data collections, the first in 2000, and the most recent in 2015.

Students from just over 30 OECD member countries and 30 non-member countries are participating in 2015. Table 2.1 lists the areas assessed. In 2018, the focus area will be reading. Students are measured on a range of other factors including attitudes and motivation. Countries participating in PISA longitudinally can compare their students' performance over time. That said, the student data is not longitudinal. A different cohort of students is tested every three years, such that data cannot be analysed at the student level. This reality clearly delineates the focus of PISA on system wide patterns.

The objectives of the programme are to develop regular, reliable and relevant indicators on student achievement, with four 'products' outlined:

a) "A set of basic indicators that will provide policy makers with a baseline profile of the knowledge, skills and competencies of students in their countries
b) A set of contextual indicators that will provide insight into how such skills relate to important demographic, social, economic and educational variables
c) Trend indicators that will become available because of the on-going cyclical nature of the data collections
d) A knowledge base that will lend itself to further focused policy analysis" (PISA, 2015b).

[1] For PISA 2015 and 2018, these are Pearson, ETS, Westat and DIPF

Table 2.1 PISA focus and country involvement 2000-2015

Assessment details	Assessment year	2000	2003	2006	2009	2012	2015
	Subjects Assessed	Reading, Mathematics, Science	Reading, Mathematics, Science, Problem Solving	Reading, Mathematics, Science	Reading, Mathematics, Science	Reading, Mathematics, Science, Creative Problem Solving, Financial literacy	Reading, Mathematics, Science, Collaborative Problem Solving, Financial literacy
	Subject focus on:	Reading	Mathematics	Science	Reading	Mathematics	Science
Economies Participating with results reported •	Albania	•			•	•	•
	Algeria						•
	Argentina	•		•	•	•	•
	Australia	•	•	•	•	•	•
	Austria	•	•	•	•	•	•
	Azerbaijan			•	•		
	Beijing-China						•
	Belgium	•	•	•	•	•	•
	Brazil	•	•	•	•	•	•
	Bulgaria1	•		•	•	•	•
	Canada	•	•	•	•	•	•
	Chile	•		•	•	•	•
	Chinese Taipei			•	•	•	•
	Colombia			•	•	•	•
	Costa Rica				•	•	•
	Croatia			•	•	•	•
	Cyprus					•	
	Czech Republic	•	•	•	•	•	•
	Denmark	•	•	•	•	•	•
	Dominican Republic						•
	Dubai-United Arab Emirates				•		
	Estonia			•	•	•	•
	Finland	•	•	•	•	•	•
	France	•	•	•	•	•	•
	Georgia				•		•
	Germany	•	•	•	•	•	•
	Greece	•	•	•	•	•	•
	Guangdong-China						•
	Himachal Pradesh-India				•		
	Hong Kong-China1	•	•	•	•	•	•
	Hungary	•	•	•	•	•	•
	Iceland	•	•	•	•	•	•
	Indonesia	•	•	•	•	•	•
	Ireland	•	•	•	•	•	•
	Israel	•		•	•	•	•
	Italy	•	•	•	•	•	•
	Japan	•	•	•	•	•	•
	Jiangsu-China						•
	Jordan			•	•	•	•
	Kazakhstan				•	•	•
	Korea, Republic of	•	•	•	•	•	•
	Kosovo						•
	Kyrgyz Republic			•	•		
	Latvia	•	•	•	•	•	•
	Lebanon						•
	Liechtenstein	•	•	•	•	•	
	Lithuania			•	•	•	•
	Luxembourg	•	•	•	•	•	•

	Macao-China		●	●	●	●	●
	Macedonia, Republic of	●					●
	Malaysia				●	●	●
	Malta				●		●
	Mauritius				●		
	Mexico	●	●	●	●	●	●
	Miranda-Venezuela				●		
	Moldova, Republic of				●		●
	Montenegro, Republic of			●	●	●	●
	Netherlands	●	●	●	●	●	●
	New Zealand	●	●	●	●	●	●
	Norway	●	●	●	●	●	●
	Panama				●		
	Peru	●			●	●	●
	Poland	●	●	●	●	●	●
	Portugal	●	●	●	●	●	●
	Puerto Rico						●
	Qatar			●	●	●	●
	Romania	●		●	●	●	●
	Russian Federation	●	●	●	●	●	●
	Serbia			●	●	●	
	Serbia and Montenegro		●				
	Shanghai-China				●	●	●
	Singapore				●	●	●
	Slovak Republic		●	●	●	●	●
	Slovenia			●	●	●	●
	Spain	●	●	●	●	●	●
	Sweden	●	●	●	●	●	●
	Switzerland	●	●	●	●	●	●
	Tamil Nadu-India				●		
	Thailand	●	●	●	●	●	●
	Trinidad and Tobago				●		●
	Tunisia		●	●	●	●	●
	Turkey		●	●	●	●	●
	United Arab Emirates				●	●	●
	United Kingdom	●	●	●	●	●	●
	United States	●	●	●	●	●	●
	Uruguay		●	●	●	●	●
	Vietnam					●	●
	Total Economies	**43**	**41**	**57**	**75**	**65**	**75 signed up**

Countries in **bold** are OECD member countries.

Table adapted from https://nces.ed.gov/surveys/pisa/countries.asp with data obtained by Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA).

2.1.2 Design and sample

Sample

The sampling techniques employed by PISA are thoroughly described in Chapter 4 of the technical report published by the OECD each year. The 2012 sample design is summarised as follows and the 2015 and 2018 sampling technique will follow the same procedures.

The testing population for PISA is 15-year-old students completing Grade 7 or higher who are enrolled full-time or part-time in educational institutes, vocational programs, foreign schools within the country, or any other related type of educational institute. No testing of students who are full-timed schooled in the home, work-place or out of the country is permitted. The 15-year-old target is slightly adapted to fit the structure of the northern hemisphere countries. The target population is set to include students from 15 years and 3 complete months to 16 years and 2 complete months. This allowed, for example, countries testing in April 2012 to test all students born in the year 1996.

The international requirement is that all countries complete all assessments within a 42-day test window. While the exact dates of the PISA test period may change, the test timing procedure allows countries to be confident that there is no disadvantage arising from different testing dates for different countries.

The sampling itself is a two-stage stratified sample design in all countries excluding Russia, which has adopted a three-stage design. The countries themselves are responsible for both the sampling and the cost of preparing the sampling system.

The first stage sampling process relates to identifying representative schools. All schools with 15-year-old students are considered to be part of the sampling frame. Schools are sampled systematically from the sampling frame, with probabilities that are proportional to a measure of size, a function of the estimated number of PISA eligible students enrolled in each school. This technique is referred to as a systematic Probability Proportional to Size (PPS) sampling method. Schools in the sampling frame are assigned to mutually exclusive groups based on school characteristics. A minimum of 150 schools per country must be tested to meet requirements. Each country's National Project Manager is encouraged to identify variables for use in the sampling process in order to reduce sampling variance.

The second stage relates to selecting students within each sampled school. For each country a Target Cluster Size (TCS) is set, which is typically 35 students per school. These students are selected with equal probability from the total number of students eligible in the school.

Quality standards relating to the PISA international target population and the school and student response rates are enforced by PISA. In brief, there are rules for within-school exclusions from the assessment for students who are intellectually disabled or functionally disabled. Whole schools can also be excluded on this basis, if the entire population of PISA eligible students would be excluded. The overall exclusion rate for a country (school-level and within-school exclusions combined) is to be kept below 5% to meet PISA requirements.

In terms of school and student response rates, there are also requirements. A response rate of 85% for initially selected schools is required. However, if the rate is above 65%, an acceptable rate can be

achieved by using replacement schools. To compensate for a sampled school which does not participate, two replacement schools are usually identified. Schools need a minimum student response rate of 50% to be considered a participating school. If a school has a student participation rate between 25% and 50%, that school is not considered as a participating school for the purposes of documenting response rates, but the student data is included and contributes to the estimates in the initial PISA international report. If the student response rate is below 25%, the school is considered a non-respondent.

An overall student response rate of 80% is required for country participation in PISA. This average is taken only from schools whose participation rates were over 50%. Weighted student response rates are used: students are weighted by the reciprocal of their sample selection probability. The field trial for PISA consists of approximately 1,500 students and the main study of approximately 6,000 students per country.

Details on the definitions of the national target population, the sampling frame, stratification variables used (by country, number of explicit strata varies from 4-104), school sampling selection, sampling frame sorting, assigning measures to each school, identifying samples schools and how to deal with PISA and national test overlap, are included in Chapter 4 of the technical report (OECD, 2014).

Test Design

PISA test questions are designed by experts from countries involved in the study; questions are constructed to represent the concepts tested. The design of the test questions is a fairly transparent process, with easy access to frameworks and example items available from PISA online (see for example 'Take the test – Sample Questions from OECD's PISA Assessments' available at http://www.oecd.org/pisa/pisaproducts/Take%20the%20test%20e%20book.pdf).

Test questions are piloted in all countries before a final test is constructed. Test booklets are then created, containing questions grouped into 'units'. Each 'unit' consists of a stimulus that may be a mixture of text, tables, figures or graphs, followed by several questions ('items' that may be in multiple-choice, short-answer or longer constructed response format. Students are given two hours to complete the assessment. In addition, students are given a separate questionnaire to gather information about their background, educational environment, families, attitudes, aspirations and learning strategies.

Tests have been paper-based from 2000 to 2012. However, from 2015 onwards, most countries will complete PISA tests by computer. The OECD has acknowledged that some countries will still require the survey in paper-based form, so computer access is required for participation.

Scorers of the tests use a detailed scoring guide to identify whether responses are correct, partly correct, or incorrect ('full credit, partial credit or no credit') for each item. Given the goal of PISA to compare the outcomes of education systems, the selection of items chosen to form this comparison should be representative of the knowledge and skills reflecting the key content areas, and be considered valuable to participant countries. The procedures for item selection are explained fully in Chapter 2 of each PISA technical report (e.g. OECD, 2009). The correction of the tests is overseen by

the National Project Manager, using a guide developed by PISA, and results are cross-checked. The item level results are then sent to PISA.

Since PISA is designed to collect internationally comparable data, the equivalence of national versions of the assessment is essential across the different languages and cultures represented by the country systems. PISA requires a double-translation and reconciliation process for finalisation of the actual assessment items. (As students in the Philippines would probably be tested in English, this review does not include information about the translation process or requirements. These details are available from the document 'Translation and Adaptation Guidelines for PISA 2012 - Doc: Tran_Adapt_Guide_PISA12'.)

The use of thoroughly prepared manuals is designed to ensure that field operations and test administration procedures are carried out uniformly across participating countries and schools. The student questionnaire takes 20-30 minutes to complete, and school principals complete a 20 minute questionnaire about their schools. Optional questionnaires are available from PISA for computer familiarity, educational career, and parent background. Countries can choose these, and/or use national questionnaires to obtain further information about their students. Student assessment data can be analysed in the context of the questionnaire information, including the identification of factors to explain differences in achievement as observed in groups of students across and within countries.

2.1.3 Use of data

PISA scores are adjusted to fit a common scale where the OECD average is 500 points (http://www.oecd.org/pisa/aboutpisa/pisafaq.htm) and the standard deviation is 100, with approximately two thirds of all OECD countries tested scoring between 400 and 600 points. Scores are located along a scale for each subject area. Level descriptions are provided, such that Level 1 includes items that require only the most basic skills to complete, with skills increasing through the succeeding levels. The score for a country is the average score of all students measured for each subject area. Rasch item response techniques are used to develop the scales.

PISA uses differential item analysis to determine whether items are 'behaving' similarly across countries. These analyses are employed to identify whether there are culturally specific item differences. The choice of scaling methods and the subsequent removal or use of items that are shown to be culturally biased is a matter of contention (see for example Kreiner & Christensen, 2013; Wuttke, 2007; and Goldstein, 2004). However, academics well versed in the Rasch scaling procedure agree that PISA uses appropriate scaling methods and deals with the challenges of constraints on sample size and differing country backgrounds adequately (Adams, Wu, & Carstensen, 2007; Adams, Berezner, Jakubowski, 2010; Grisay, de Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007; Le, 2006a; Le, 2007), using procedures outlined by Mislevy and others (Mislevy, Beaton, Kaplan, & Sheehan 1992). Differences due to gender are also analysed, and items showing bias are subjected to similar scrutiny (Le, 2006b; 2009).

PISA uses a pool of assessment items. If some items do not work well statistically in a particular country (items showing bias) they may be removed from the analysis entirely or from the particular country's analysis for the purpose of student score generation. The items may remain in the assessment in other countries, where there were no statistical issues or bias. In addition, students

within the one country do not complete exactly the same assessment. Sampling of items within a country's sample population allows an analysis of student proficiency without having to assess every student on every item. This is a fundamental advantage of the Rasch measurement process. Test booklets are linked psychometrically and scaled so that each student can be placed on the same scale, developed empirically for each subject area. More information about the scaling process used for PISA is outlined in the technical booklets.

2.1.4 Implementation considerations

PISA is financed exclusively by participating countries. Each country is responsible for paying for the use of the assessment and the cost of national implementation. In 2015 the cost for new participants was EUR 182,000 payable over four years at 45,500 per year from 2013 to 2015 inclusive.

In addition to the cost of the assessment, each country is responsible for the costs of their national implementation, including:

- Employment of expert sampling statisticians to take responsibility for drawing a representative sample of schools and students as outlined in Section 2.1.2 (the field trial will consist of approximately 1,500 students and the main study approximately 6,000 students, with at least 150 schools sampled)
- Resourcing of authorities to recruit schools to participate and administer the tests
- Provision of personnel to prepare assessment booklets
- Provision of personnel to process returned booklets and to score tests, including open-ended test items
- Contribution to international overhead costs.

Participating countries appoint a National Project Manager (NPM) to oversee implementation. The NPM works with the OECD contractor on all issues relating to the implementation of PISA in their country. The NPM should have a university degree and previous experience in planning, organising and administering large-scale surveys, project management experience, excellent written and oral communication skills in English, and knowledge of educational systems. The NPM is involved in development and review of PISA reports and documentation and attends meetings with other NPMs. For example for 2015, there will be a total of six international meetings for NPMs. Meetings will follow the approximate timeline for 2018 survey participants: September 2015, March 2016, February 2017, October 2017, February 2018 and October 2018.

Participant countries can nominate a representative for the PISA Governing Board. This board is responsible for specifying the policy priorities and standards for development of indicators, establishment of the assessment instruments and the reporting of the results. The board meets twice a year, in March/April and in October/November. Representation on the board is optional for non-member OECD countries. However, participation is arguably of great benefit to any country involved in the PISA survey.

The first National Project Meeting for the 2018 round will probably be held in September 2015.

The PISA Governing Board approves membership according to certain criteria. In brief, participants must have the technical expertise to administer an international assessment and must be able to meet the full cost of participation.

Applications to participate in PISA are formally submitted to the OECD, with confirmation of an intention to contribute to the international overhead costs. Letters should be addressed to: **Mr. Andreas Schleicher, Directorate for Education (juliet.evans@oecd.org).** More information regarding participation can be viewed at http://www.oecd.org/pisa/aboutpisa/howtojoinpisa.htm.

2.2 Programme for International Student Assessment for Development

2.2.1 Summary and aims

The OECD and partners[2] launched a three-year 'PISA for Development' initiative in 2013, the aim of which was to identify how PISA could best support evidence-based policy making in emerging economies and developing nations. The premise behind the initiative was to make PISA even more relevant for a broader set of countries and contribute to the United Nations-led global learning goals (otherwise known as Millennium Development Goals-MDGs), which were developmental objectives to be achieved by 2015. To meet these objectives, the following steps were undertaken (PISA, 2013);

1. "Developing contextual questionnaires and data-collection instruments that better capture diverse situations in emerging and developing countries. This will allow for a deeper understanding of how certain factors – such as the socio-economic background of students or the learning environment in classrooms – are associated with learning outcomes in different contexts.
2. Adjusting the PISA test instruments so that they are sensitive to a wider range of performance levels. While there are undoubtedly high performers in all countries, a number of 15-year-old students in developing countries can be expected to perform at lower levels of proficiency. Enhanced test instruments will better capture performance differences among these students, while maintaining the comparability of a country's results on the international PISA scales.
3. Establishing methods and approaches to include out-of-school students in the PISA assessment. Though much progress has been made in increasing access to education around the world, over 60 million children of primary-school age and over 70 million children of lower-secondary-school age remain out of school. Conducting PISA only among enrolled students would provide unrepresentative results and could encourage countries to exclude potential low performers from schools."

With these efforts, the OECD hopes that more countries will have the opportunity to use PISA to set national learning targets, monitor progress towards those targets and to analyse the factors that affect student outcomes among poor and marginalised populations.

[2] Partner countries described in section 2.2.2

2.2.2 Partner countries

The countries participating in the PISA for Development pilot are Ecuador, Guatemala, Senegal, and Zambia, negotiations are underway with Cambodia and Paraguay (PISA, 2015). The national project managers (NPMs) from these participating countries implement PISA for Development at the national level, subject to the agreed administration procedures and are responsible for ensuring that the implementation is of high quality, and for verifying and validating the survey results, analyses reports and publications.

The results from the PISA for Development pilot project have not yet been released, but the use of the results are hoped to retain a focus on access and equity at primary and secondary education levels, with a meaningful focus on learning.

2.3 Trends in International Mathematics and Science Study (TIMSS)

2.3.1 Summary and aims

The Trends in International Mathematics and Science Study (TIMSS) is an international assessment of mathematics and science that was first conducted in 1995. TIMSS measures the science and mathematics ability of Grade 4 and Grade 8 students. The International Association for the Evaluation of Educational Achievement (IEA), an independent international cooperative of national research institutes and government agencies, runs the assessment every four years.

TIMSS is designed to provide information that will assist countries to monitor and evaluate the success of their mathematics and science education across time and across grades. The intention of this assessment is to improve teaching and learning of mathematics and science by providing information about student achievement in relation to different types of curricula, instructional practices, and schools. TIMSS also aspires to inform policy in the participating countries around the world (Martin & Mullis, 2006).

The number of countries participating in different cycles of this assessment is presented in Table 2.2. More detailed lists of education systems (countries, states and benchmarking participants) are presented in Tables 2.3 to 2.6.

Table 2.2 Number of countries participating in different cycles of TIMSS

Assessment Year	4th grade	8th grade
2015	49 Countries + 5 benchmarking entities	42 Countries + 4 benchmarking entities
2011	52 Education Systems	45 Education Systems
2007	44 Education Systems	57 Education Systems
2003	26 Education Systems	48 Education Systems
1999	-	38 Education Systems
1995	41 Education Systems across 5 grades	

The 2015 cycle of TIMSS will report overall achievement in science and mathematics as well as results across four international benchmarks (advanced, high, medium, and low), by major content domains (ie. number, algebra, and geometry in mathematics; earth science, biology, and chemistry in science), and by cognitive domains (knowing, applying, and reasoning). The study will also collect information regarding curriculum and curriculum implementation, instructional practices, and school resources.

In addition to the regular TIMSS assessment, the IEA administers a less difficult assessment, TIMSS Numeracy, and a more difficult assessment, TIMSS Advanced. These are outlined in Sections 2.3 and 2.4 below.

Table 2.3 TIMSS Participating Countries (1995 – 2015)

Assessment Year	**1995**		**1999**	**2003**		**2007**		**2011**		**2015**	
Education Systems	**4th grade**	**8th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**
Algeria						●	●				
Argentina		○			○						
Armenia				●	●	●	●	●	●	●	●
Australia	●	●	●	●	●	●	●	●	●	●	●
Austria	●	●				●		●			
Azerbaijan								●			
Bahrain					●		●	●	●	●	●
Belgium (Flemish)-BEL		●	●	●	●			●		●	
Belgium (French)-BEL		●									
Bosnia and Herzegovina							●				
Botswana					●		●				●
Bulgaria		●	●		●		●			●	
Canada	●	●	●							●	●
Chile			●		●			●	●	●	●
Chinese Taipei			●	●	●	●	●	●	●	●	●
Colombia		●				●	●				
Croatia								●		●	
Cyprus	●	●	●	●	●		●			●	
Czech Republic	●	●	●			●	●	●		●	
Denmark		●				●		●		●	
Egypt					●		●				●
El Salvador						●	●				
England-GBR	●	●	●	●	○	●	●	●	●	●	●
Estonia					●						
Finland		●	●					●	●	●	
France										●	
Georgia						●	●	●	●	●	●
Germany		●				●		●		●	
Ghana					●		●		●		
Greece	●	●									
Hong Kong	●	●	●	●	●	●	●	●	●	●	●

Hungary	●	●	●	●	●	●	●	●	●	●	●
Iceland	●	●									
Indonesia	○	○	●		●		●		●	●	
Iran	●	●	●	●	●	●	●	●	●	●	●
Ireland	●	●						●		●	●
Israel	●	●	●		●		●		●		●
Italy	○	○	●	●	●	●	●	●	●	●	●
Japan	●	●	●	●	●	●	●	●	●	●	●
Jordan			●		●		●		●		●
Kazakhstan						●		●	●	●	●
Korea Republic	●	●	●		●		●	●	●	●	●
Kuwait	●	●				●	●	●		●	●
Latvia	●	●	●	●	●	●					
Lebanon					●		●		●		●
Lithuania		●	●	●	●	●	●	●	●	●	●
Macedonia, Republic of			●		●				●		
Malaysia			●		●		●		●		●
Malta							●	●			●
Mexico	○	○									
Moldova			●	●	●						
Mongolia						○	○				
Morocco			●	●	●	●	○	●	●	●	●
Netherlands	●	●	●	●	●	●		●		●	
New Zealand	●	●	●	●	●	●		●	●	●	●
Northern Ireland-GBR								●		●	
Norway	●	●		●	●	●	●	●	●	●	●
Oman							●	●	●	●	●
Palestinian Nat'l Auth.					●		●		●		
Philippines		○	●	●	●						
Poland								●		●	
Portugal	●	●						●		●	
Qatar						●	●	●	●	●	●
Romania		●	●		●		●	●	●		
Russian Federation		●	●	●	●	●	●	●	●	●	●
Saudi Arabia					●		●	●	●	●	●
Scotland-GBR	●	●		●	●	●	●				
Serbia					●		●	●		●	
Singapore	●	●	●	●	●	●	●	●	●	●	●
Slovak Republic		●	●		●	●		●		●	
Slovenia	●	●	●	●	●	●	●	●	●	●	●
South Africa		●	●		●						●
Spain		●						●		●	
Sweden		●			●	●	●	●	●	●	●
Switzerland		●									
Syrian Arab R							●		●		
Thailand	●	●	●				●	●	●		●
Tunisia			●	●	●	●	●	●	●		
Turkey			●				●	●	●	●	●

Ukraine						●	●		●		
UAE								●	●	●	●
USA	●	●	●	●	●	●	●	●	●	●	●
Yemen				○		●		●			
Total	**29**	**46**	**38**	**26**	**47**	**37**	**50**	**50**	**42**	**48**	**40**

Table 2.4 US benchmarking states[1]

Assessment Year	**1995**		**1999**	**2003**		**2007**		**2011**		**2015**	
Education Systems	**4th grade**	**8th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**
Alabama									●		
California									●		
Colorado	●								●		
Connecticut			●						●		
Florida								●	●	●	●
Idaho			●								
Illinois		●	●								
Indiana			●	●	●				●		
Maryland			●								
Massachusetts			●			●	●		●		
Michigan			●								
Minnesota	●	●				●	●		●		
Missouri[2]		●	●								
N Carolina-USA			●					●	●		
Oregon[2]		●	●								
Pennsylvania-			●								
South Carolina-			●								
Texas			●								
Total	**2**	**4**	**13**	**1**	**1**	**2**	**2**	**2**	**9**	**1**	**1**

Table 2.5 Other benchmarking participants

Assessment Year	**1995**		**1999**	**2003**		**2007**		**2011**		**2015**	
Education Systems	**4th grade**	**8th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**
Abu Dhabi-UAE								●	●	●	●
Alberta-CAN	●	●	●			●		●	●		
Basque Country-ESP					●		●				
British Columbia-CAN			●			●	●				
Buenos Aires-ARG										●	●
Dubai-UAE						●	●	●	●	●	●
Ontario-CAN	●	●	●	●	●	●	●	●	●	●	●
Quebec-CAN	●	●	●	●	●	●	●	●	●	●	●
Total	**3**	**3**	**4**	**2**	**3**	**5**	**5**	**5**	**5**	**5**	**5**

Table 2.6 Off-grade participants

Assessment Year	1995		1999	2003		2007		2011		2015	
Education Systems	**4th grade**	**8th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**	**4th grade**	**8th grade**
Botswana[3,4]								●	●		●
Honduras[3,4]								●	●		
South Africa[4]									●		●
Yemen[3]								●			
Total								**3**	**3**		**2**

● = Indicates participation in particular assessment with results reported or forthcoming.
○ = Indicates participation in particular assessment but results either not reported or reported separately, typically due to sampling, response rates, or other procedural problems with the data.
[1] U.S. state decisions regarding TIMSS participation in 2015 are not yet finalized.
[2] Missouri-USA and Oregon-USA participated in the 1997 TIMSS Benchmarking Study, which administered the TIMSS 1995 assessment to 8th-grade students in 1997.
[3] Administered the TIMSS 4th-grade assessment to 6th-grade students in 2011.
[4] Administered the TIMSS 8th-grade assessment to 9th-grade students in 2011. Botswana and South Africa plan to administer the TIMSS 8th-grade assessment to 9th-grade students in 2015.
NOTE: OECD member countries are bolded. Subnational education systems are italicized.
SOURCE: Table adapted from https://nces.ed.gov/TIMSS/countries.asp with data obtained by International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS).

As summarised by Thomson et al. (2012), TIMSS offers countries an opportunity to:

a) have comprehensive and internationally comparable data about the mathematics and science concepts, processes and attitudes have been learned by Grade 4 and Grade 8 students;
b) assess progress internationally in mathematics and science learning across time for Grade 4 and Grade 8 students;
c) identify aspects of growth in mathematical and scientific knowledge and skills from Grade 4 to 8;
d) monitor the relative effectiveness of teaching and learning of mathematics and science at Grade 4 compared to Grade 8, since the cohort of Grade 4 students is assessed again in Grade 8;
e) understand the contexts in which students learn best - TIMSS enables international comparisons according to key policy variables in curriculum, instruction and resources that result in higher levels of student achievement;
f) use TIMSS to address internal policy issues - within countries, for example, TIMSS provides an opportunity to examine the performance of population subgroups and address equity concerns;
g) allow countries to add questions of national importance (national options) as part of their data collection effort. (p. 2)

2.3.2 Sample and design

Sample

The international target population for TIMSS consists of Grade 4 and 8 students. Countries can also choose to administer some tools to a grade level proximate to that intended.

TIMSS employs school and classroom sampling techniques that require the assessment of a selection of students from a selection of schools. Countries can assess Grade 4 or Grade 8 or both. TIMSS implements a two-stage random sample design in which a sample of schools is drawn in the first stage and one or more intact classes of students are selected from each of the sampled schools in the second stage. Random-start fixed-interval systematic sampling is employed to draw the school sample, in which each school is selected with probability proportional to its size (PPS) (Joncas & Foy, 2012). Afterwards, intact classes are selected from the sampled schools. The assessment for TIMSS pays particular attention to curricular and instructional experiences of students. Since these are typically organized on a classroom basis, intact classes of students are sampled rather than individuals from across the grade level or of a certain age.

Participating countries need a plan to define their national target population and to apply the TIMSS sampling methods to achieve a nationally representative sample of schools and students. The National Research Coordinator (NRC) and TIMSS sampling experts collaborate to develop and implement the national sampling plan. Statistics Canada is responsible for advising the National Research Coordinator on all sampling matters and for ensuring the plan conforms to the TIMSS standards.

The TIMSS guidelines state that a minimum of 150 schools must be sampled per grade, and a minimum of 4,000 students must be sampled. Countries are allowed to draw larger samples of schools and students, and some do. For example, in TIMSS 2011 in the USA, 369 schools and 12,569 students participated in the Grade 4 assessment, and 501 schools and 10,477 students participated in the Grade 8 assessment.

For a country's data to be included in the international database, a minimum participation rate of 50% of schools from the original sample of schools is required. The target response rate for classrooms is 95% and the target student response rate is 85%. Countries are allowed to use substitute schools, selected during the sampling process, if some schools originally sampled need to be excluded for compelling reasons. Substitute schools are required to be in the same identified stratum and are identified as the two schools neighbouring the sampled school in the frame. Substitute schools can be selected only after the minimum participation target of 50% has been reached from the original sample of schools (Joncas & Foy, 2011).

Test design

The TIMSS assessment of students' achievement comprises written tests in mathematics and science with multiple-choice and constructed-response items, together with sets of questionnaires that gather information on the educational and social contexts of the students, their teachers and their schools. TIMSS pays particular attention to covering the breadth and richness of mathematics and science, which means that many more questions are required for the assessment than can be

answered by any one student in the testing time provided. Accordingly, TIMSS uses a matrix-sampling approach and packages the entire assessment pool of mathematics and science items at each grade level into booklet sets (in TIMSS 2011 there were 14 different booklets, each carrying items on mathematics and science). Each student needs to complete one booklet only. The booklets are constructed in such a way that their difficulty levels are similar and each booklet is distributed to groups of students with approximately equivalent levels of ability. Each has two blocks of mathematics items and two blocks of science items. These blocks contain approximately 10-14 items at Grade 4 (72 minutes) and 12-18 items at the Grade 8 (90 minutes). In addition, 30 minutes are allocated for the student questionnaire at each grade level. More details regarding the booklet preparation and test construction can be found in the TIMSS 2011 Assessment Frameworks (available at http://timssandpirls.bc.edu/timss2011/frameworks.html) and the TIMSS 2015 Assessment Frameworks (available at http://timssandpirls.bc.edu/timss2015/frameworks.html).

While developing items for TIMSS, consideration is given to the existing pool of trend items, which is used to compare the results with previous years' assessment. The mathematics and science trend items are mapped onto the content and cognitive domains outlined in the Assessment Framework. In addition, new items are developed by experts through a rigorous process designed to ensure that they complement the existing set of items. More details on how the items are developed and contextualized for use in different countries can be found in the assessment frameworks mentioned above.

The NRCs for participating countries are given in-depth training on how to score the items. The constructed response items are scored following a guide that describes the essential features of appropriate and complete responses. The guide focuses on evidence of the type of behaviour the items assess and describes characteristics of partially correct and completely correct responses (Martin, Mullis, & Foy, 2013).

Countries participating in TIMSS aim for a sample of at least 4,500 students to ensure that there are sufficient respondents for each item. The student booklets are distributed so that approximately the same number of students will respond to each booklet. TIMSS uses item response theory scaling methods to gain an overall picture of students' achievement from the combined responses of individual students to the booklets they are given.

The student questionnaire collects demographic information and asks students about various aspects of their home and school life. It includes questions about self-perception and attitude toward learning mathematics and science. Teachers of the assessed classes complete a questionnaire that collects data such as their education, professional development, and experience in teaching. In addition, information is requested on characteristics of the class, instructional time, materials, activities for teaching mathematics and science, etc. Principals of sampled schools complete a school questionnaire on student demographic characteristics, availability of resources, types of programs, and environment for learning in the school. Finally, the NRC in each participating country is responsible for completing a curriculum questionnaire. This questionnaire primarily elicits information about the organisation and content of the mathematics and science curriculum (Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009).

In 2019, the TIMSS assessment will be available in an electronic form, to be called eTIMSS, as well as in the usual paper and pencil form. The paper and pencil test will remain available to ensure than IEA does not preclude any country from participation.

2.3.3 Use of data

As mentioned, each student participating in TIMSS responds to only a subset of the TIMSS mathematics and science item pool. Due to the complexities of the data collection, and the necessity, for analysis and reporting purposes, of obtaining student scores across entire assessments, TIMSS depends on item response theory (IRT) scaling to describe student achievement. The TIMSS scaling approach uses multiple imputations or 'plausible values' methods to obtain scores in mathematics and science. Reliability is enhanced by a process called 'conditioning', in which student responses to test items are combined with their background information (Foy, Brossman, & Galia, 2013).

Analysis of TIMSS assessment data uses three IRT models, all of which are 'latent variable' models. TIMSS multiple choice items are scored dichotomously (correct/incorrect) and, depending on the scoring guide, constructed response items are scored either dichotomously (correct/incorrect) or polytomously for partly correct (partial credit). Accordingly, a three-parameter IRT model is used for multiple choice items, a two-parameter model is used for constructed response items scored dichotomously, and a partial credit model is used for constructed response items scored polytomously (Foy, Brossman, & Galia, 2013).

Corresponding to the international means and standard deviations across all the countries that participated in TIMSS 1995, the TIMSS mathematics and science achievement scales were established in 1995 to have a scale average of 500 and a standard deviation of 100. Results from all subsequent cycles of TIMSS have been placed on the same scale. This common metric gives the participating countries an opportunity to compare their Grade 4 and 8 students' progress on mathematics and science from one time to the next (Mullis et al., 2009).

TIMSS provides internationally comparable information across participating countries. According to IEA, the uses of TIMSS data include:

i) Monitoring the effectiveness of education systems in a global context: TIMSS provides comprehensive data about what mathematics and science concepts, processes and attitudes students have learned by Grades 4 and 8. The data can be compared to the global context and gives an understanding of individual countries' comparative positions.
ii) Evaluating progress in educational achievement: TIMSS gives the opportunity to evaluate progress from both national and international perspectives. If achievements are lower than expected, countries can take steps to stimulate improvement.
iii) Working towards closing gaps between achievements: TIMSS data can be used to achieve equity among different ethnic, social or regional groups. Some countries have made specific efforts to reduce achievement disparities between groups of students after analysis of results. Note that it is possible for countries to add questions of national importance to questionnaires (national options) or over-sample particular groups of students as part of the data collection effort to get an insight into the educational needs of particular groups.

iv) Examining educational achievement and growth from primary through secondary school: TIMSS data allows the participating countries to measure the growth of their students in mathematics and science learning throughout the schooling years. Moreover, TIMSS data frameworks and released items have served as a basis for curriculum reform and designing teacher education in almost every participating country. This is particularly applicable for the areas of problem solving, reasoning, and inquiry.

v) Using the international database for TIMSS to research the factors associated with high achievement: TIMSS has an international database that contains all the data collected through different cycles of the assessment in different countries. Most of the participating countries have conducted research using this database to understand the school and classroom contexts in which students learn mathematics and science best, including curricular variation, resources, administrative and instructional policies.

More details regarding the value of participating in TIMSS and documents on how previous TIMSS assessment data have been used by participating countries can be found at http://timss.bc.edu/timss2015/participate.html.

2.3.4 Implementation considerations

Both IEA member and non-member countries are welcome to join IEA studies (including TIMSS). Participating countries are expected to establish their own national centre and appoint an NRC, as well as a national committee consisting of experts in the curriculum and policy-making domains and in the technical aspects of implementing the study. The committee must be available for consultation throughout the duration of the project.

Each participating country is also required to cover the costs of the study at the national level (including the costs that will allow their NRC to attend study meetings) and to contribute to the costs of coordinating the study internationally. The amount of funding requested of participating countries depends on the scope of the study and the availability of funds from other national or international agencies. The secretariat establishes these costs in close cooperation with the international study centre and IEA membership.

The next TIMSS assessment will be held in 2019. If current timeline patterns are followed, it is likely that participation decisions will need to be finalised by the end of 2016. A brief overview of the schedule of TIMSS 2015 provides an indication of what might be expected in the next phase. The first NRC meeting for TIMSS 2015 was held in February 2013. Framework and instrument development work was carried out throughout 2013 and the field test was conducted in March–April 2014. The data collection of the main survey took place in October–December 2014 (southern hemisphere countries) and March–June 2015 (northern hemisphere countries). The international reports are scheduled to be released in December 2016, followed by the international database and user guide in February 2017.

Funding of TIMSS

The TIMSS assessment is funded by finance from the IEA and fees from participating countries. In addition, some funding is obtained from the National Centre for Education Statistics of the United States Department of Education. Fees for participation are assessed in two currencies and on a

yearly basis for each of the four years of the project. The TIMSS 2015 participation fee is USD 25,000 (EUR 25,000) per year for one grade. The yearly participation fee for TIMSS Advanced 2015 is USD 37,500 (EUR 37,500). For countries participating in TIMSS 2015 at two grades or TIMSS 2015 and TIMSS Advanced 2015 together, there is a reduction in fees.

2.4 TIMSS Numeracy

In 2015, the IEA is introducing a less challenging mathematics assessment: TIMSS Numeracy. TIMSS Numeracy aims to address the needs of the global education community and its efforts to work towards universal learning for all children. Ways to measure progress towards learning goals are needed as the debate shifts from *access* for all to *learning* for all. This assessment aims to help countries and international organizations measure and improve learning outcomes for children and youth worldwide.

TIMSS Numeracy assesses fundamental mathematical knowledge, procedures, and problem solving strategies that are prerequisites for success on TIMSS Fourth Grade. The test items are similar to TIMSS Fourth Grade items, but the numbers are simpler and the procedures are more straightforward. The assessment is designed to test mathematical knowledge and skill towards the end of the primary or elementary school cycle and can be administered to students in Grades 4-6.

TIMSS Numeracy 2015 comprises 10 blocks of items, each containing 10-15 items. Since 2015 is the first time the assessment has been administered, all the items are newly developed, although the item blocks were developed following the same guidelines as TIMSS Fourth Grade. The question types (multiple choice and constructed response items) and scoring procedures are also the same. Like students undertaking TIMSS Fourth Grade, students undertaking TIMSS Numeracy are expected to spend an average of 18 minutes on each item block. The 10 item blocks are distributed across five student achievement booklets, with each booklet containing four blocks of numeracy items. Each block of items appears in two booklets, enabling linking across booklets. The assessment time for each booklet (i.e. assessment time for each student) is 72 minutes. An additional 30 minutes is allocated for the student questionnaire.

Since 2015 is the maiden administration of TIMSS Numeracy, limited information is available and no data are currently available on the number of participating countries.

2.5 TIMSS Advanced

2.4.1 Summary and aims

TIMSS Advanced assesses achievement in advanced mathematics and physics among students completing secondary school and entering tertiary education. It was administered in 1995 and 2008 to students in the final year of secondary school and can be administered in 2015 to students in either the final year of secondary school or the first year of tertiary education immediately following their graduation from secondary school. In addition, data is collected on curriculum emphasis, technology use, and teacher preparation and training.

Table 2.7 TIMSS Advanced Participating Countries (1995 – 2015)

Education system	1995	2008	2015
	Last-year Secondary School	Last-year Secondary School	Last-year Secondary School / First-year Tertiary Education
Armenia		●	
Australia	●		
Austria	●		
Canada	●		
Cyprus	●		
Czech Republic	●		
Denmark	●		
France	●		●
Germany	●		
Greece	●		
Iran, Islamic Republic of		●	
Israel	●		
Italy	●	●	●
Latvia[1]	●		
Lebanon		●	
Lithuania[2]	●		
Netherlands		●	
Norway[1]	●	●	●
Philippines		●	
Portugal			●
Russian Federation	●	●	●
Slovenia	●	●	●
Sweden	●	●	●
Switzerland	●		
United States	●		●
Total	**19**	**10**	**8**

● = Indicates participation in particular assessment with results reported or forthcoming.

[1] Administered physics but not advanced mathematics in 1995.

[2] Administered advanced mathematics but not physics in 1995.

SOURCE: Table adapted from https://nces.ed.gov/TIMSS/countries_advanced.asp with data obtained by International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS).

Many countries consider it important to ensure that capable secondary school students are given preparation in advanced mathematics and science to prepare them for entry to challenging university studies that will qualify them for careers in the 'STEM' fields of science, technology, engineering and mathematics (Mullis, 2014). These students are expected to be the scientists and engineers of the future and are anticipated to drive innovation and technological development in all sectors of their countries' economies. TIMSS Advanced is an assessment that has as its focus this group of students, assessing their advanced mathematics and physics performance and providing participating countries with information on:

- The number of students and the proportion of the student population participating in advanced mathematics and physics study;
- The achievement of these students on international benchmarks (advanced, high, or intermediate); and
- A rich set of contextual data (e.g. curricula, teaching-learning strategies, technology use, teacher preparation and training, etc.) that can be used to guide education reform and policy planning in STEM fields.
  (Mullis, 2014)

The primary motivation for countries participating in TIMMS Advanced is to gather data that will help them understand how well they are preparing a future generation of scientists and engineers. Another benefit of participation for countries that also participate at Grades 4 and 8 TIMSS is the provision of a comprehensive set of data at three points across their country's education system. A list of participating countries is provided in Table 2.7.

2.4.2 Design and sample

Sample

The target population for advanced mathematics is students who are in the final year of secondary schooling and have taken courses in advanced mathematics. Similarly the target population for physics is students in the final year of secondary schooling who have taken advanced physics courses. In countries with a tracked educational system, student eligibility will be determined by the track to which a student belongs. Deciding which mathematics and physics courses are advanced in order to define the target population lies with the participating countries. But generally the courses included will be those taken by the most advanced students who are planning to take further studies in physics or mathematics at the university level. In addition, courses that will define the target population must cover most, if not all, of the advanced mathematics and physics topics outlined in the TIMSS Advanced Assessment Frameworks (TA08 Technical Report). Depending on the courses chosen, students may belong to the advanced mathematics population, the advanced physics population or both. Students who belong to both populations are randomly selected for either advanced mathematics or advanced physics assessment (Garden et al., 2006; Mullis & Martin, 2014).

TIMSS Advanced uses a uniform sample design that can be adapted to the specific sampling requirements of individual countries. The sampling technique involves a two-stage stratified cluster sampling, where the first stage consists of schools, and the second stage consists of one or more intact classrooms from the list of eligible classes in the sampled schools. In countries where the number of schools in the population is much higher than the number required in the sample, a systematic-probability-proportional-to-size (PPS) sampling method is used. Followed by the second sampling stage where classes are selected within a school, this method is often referred to as systematic two-stage PPS sampling. On the other hand, in countries where number of schools to be sampled from is relatively small, schools are selected with equal probabilities. Classes are sampled within selected schools using a random systematic sampling in all countries.

In TIMSS Advanced 2008, participating countries sampled 120 schools and one classroom from each of them. In order to tailor the basic design to its particular situation, each country worked closely with Statistics Canada ensuring the most effective coverage of the target population as well as maximising the comparability across different countries (Arora, Foy, Martin, & Mullis, 2009). The minimum sample sizes required for TIMSS Advanced 2008 were set at 2000 tested students for mathematics and 2000 tested students for physics, selected from a minimum of 120 schools. As these were the minima, most countries targeted a larger number of schools and students as a safeguard against no responses. More details on the sampling techniques used by TIMSS Advanced can be found in the TIMSS Advanced 2008 Technical Report (available at http://timss.bc.edu/timss_advanced/downloads/TA08_Technical_Report.pdf). Countries participating in TIMSS Advanced 2015 are required to sample a minimum of 3600 advanced mathematics students and the same number of physics students.

Test Design

The TIMSS Advanced assessment comprises written tests in advanced mathematics and physics together with sets of questionnaires to gather information on educational and social contexts. To ensure thorough coverage of the assessment topic while maintaining a reasonable burden on the students' time, TIMSS Advanced uses a matrix-sampling approach where pools of achievement items in advanced mathematics and physics are assembled into set of assessment booklets – with each participating student completing one booklet only. TIMSS Advanced 2015 has 12 booklets (6 physics and 6 advanced mathematics) whereas TIMSS Advanced 2008 had 8 booklets (4 for physics and 4 for advanced mathematics). To prepare the booklets, TIMSS Advanced groups items into a series of blocks where each item block consists of approximately 10 items and requires 30 minutes of assessment time. Each item-block consists of two item formats (multiple choice and constructed response) and on an average provides about 15 points. The exact number of score points and the exact distribution of question types per block vary to some extent.

The distribution of items across content and cognitive domains within each block matches the distribution across the overall item pool as far as possible. Any given item appears in two booklets providing a mechanism for linking student responses from different booklets. To ensure that the groups of students completing each booklet are approximately equivalent in terms of ability, booklets are distributed randomly among students in participating classrooms. Furthermore, in each cycle of assessment some item blocks are retained for use in future cycles which helps in measuring trends across different cycles. The rest of the items are usually available in the public domain.

Two item formats are used in TIMSS Advanced: multiple-choice and constructed response. At least half of the total number of score points comes from multiple choice items, where each item is worth one point. The remaining score points come from constructed response items, which can be worth one or two points depending on the task and the skills necessary in completing them. Partial credits are allowed in constructed response items.

As approximately one-third of the TIMSS Advanced assessment items are constructed response items, scoring them in a reliable manner is critical to the quality of the results. Detailed scoring guides are provided, with extensive training in their use, and there is monitoring of the quality of

scoring. An international session is conducted with the NRCs of TIMSS Advanced to train them how to score the constructed response items . Moreover, to establish the reliability of scoring within each country, two different scorers independently score 25 percent of all student responses.

An important aspect of TIMSS Advanced is studying the educational contexts within which students learn advanced mathematics and physics. TIMSS Advanced administers a series of questionnaires for curriculum specialists, school principals, mathematics and physics teachers, and the students themselves.

2.4.3 Use of data

Each student participating in TIMSS Advanced responds to a subset of either the advanced mathematics or physics item pool. Considering the complexities involved in data collection and distribution of items across booklets, TIMSS Advanced uses IRT scaling methods. Scores from both advanced mathematics and physics are scaled on a scale that has a mean score of 500 and a standard deviation of 100. More details on the technical aspects of developing and using this scale are available in the TIMSS Advanced 2008 International Report (available at http://timssandpirls.bc.edu/timss_advanced/ir.html).

2.4.4 Implementation considerations

These are the same as for the main TIMSS. See Section 2.2.4.

Funding of TIMSS

The TIMSS assessments are funded by funds from the IEA and fees from participating countries. In addition, they get from the National Centre for Education Statistics of the United States Department of Education. The yearly participation fee for TIMSS Advanced 2015 is USD 37,500 (EUR 37,500). For countries participating in TIMSS 2015 at two grades or TIMSS 2015 and TIMSS Advanced 2015 together, there is a reduction in fees.

2.6 Progress in International Reading Literacy Study (PIRLS)

2.6.1 Summary and aims

Progress in International Reading Literacy Study (PIRLS) is a reading comprehension assessment conducted at five-year intervals by the International Association for the Evaluation of Educational Achievement (IEA). The next PIRLS assessment, in 2016, is the fourth cycle of this assessment with previous assessments taking place in 2001, 2006 and 2011. PIRLS is a collaborative effort of the participating countries and IEA. It is directed by the TIMSS & PIRLS International Study Centre located at Boston College, in cooperation with the IEA Secretariat in Amsterdam and IEA's Data Processing and Research Centre in Hamburg.

PIRLS aims to provide internationally comparable data on how well children read after four years of primary school. It collects extensive information about home support for literacy, curriculum and curriculum implementation, instructional practices, and school resources in each participating

country. PIRLS assesses reading literacy of students in their fourth year of formal schooling. The fourth year of school is chosen because this is considered to be an important transition point when students have learned to read and are reading to learn (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009). In many countries this is when students start to have separate classes for different subjects (e.g. mathematics and science) (Martin, Mullis, & Foy, 2015). Considering the linguistic and cognitive demands of reading, PIRLS aspires to avoid assessing very young children and recommends countries assess the next higher grade if the average age of fourth grade students at the time of testing is less than 9.5 years. Fourth grade students are beginning to engage in reading for two main purposes: reading for literacy experience and reading for acquiring and using information; these areas are assessed in PIRLS.

Countries choose to participate in PIRLS to gather data that can inform educational policy and practice by providing an international perspective on teaching and learning of reading literacy. Mullis et al. (2009) highlight the definition of reading literacy that PIRLS follows.

> "For PIRLS reading literacy is defined as the ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment" (p. 11)

A brief overview of the number of participants in different cycles of this assessment can be found in Table 2.8. In addition, a more detailed list of education systems (countries, states, and benchmarking participants) is available in Tables 2.9 through to 2.11.

Table 2.8 Number of participants in different cycles of PIRLS

Assessment Year	Education Systems	Benchmarking Participants	Off-grade Participants	Total
2011	45	8	5	**58**
2006	40	5	2	**47**
2001	34	2	1	**37**

* A full list of participant countries for PIRLS 2016 is not yet available.

Table 2.9 List of education systems participating in PIRLS (2001 – 2011)

Education system	Assessment year		
	2001	**2006**	**2011**
Argentina	●		
Australia			●
Austria		●	●
Azerbaijan			●
Belgium (Flemish)-BEL		●	
Belgium (French)-BEL		●	●
Belize	●		
Bulgaria	●	●	●
Canada			●
Chinese Taipei		●	●
Colombia	●		●

Croatia			●
Cyprus	●		
Czech Republic	●		●
Denmark		●	●
England-GBR	●	●	●
Finland			●
France	●	●	●
Georgia		●	●
Germany	●	●	●
Greece	●		
Hong Kong-CHN	●	●	●
Hungary	●	●	●
Iceland[1]	●	●	
Indonesia		●	●
Iran, Islamic Republic of	●	●	●
Ireland			●
Israel[2]	●	●	●
Italy	●	●	●
Kuwait	●	●	
Latvia	●	●	
Lithuania	●	●	●
Luxembourg		●	
Macedonia, Republic of	●	●	
Malta			●
Moldova, Republic of	●	●	
Morocco[3]	●	●	●
Netherlands	●	●	●
New Zealand	●	●	●
Northern Ireland-GBR			●
Norway[1]	●	●	●
Oman			●
Poland		●	●
Portugal			●
Qatar[2]		●	●
Romania	●	●	●
Russian Federation	●	●	●
Saudi Arabia			●
Scotland-GBR	●	●	
Singapore	●	●	●
Slovak Republic	●	●	●
Slovenia	●	●	●
South Africa		●	
Spain		●	●
Sweden[4]	●	●	●
Trinidad and Tobago		●	●
Turkey	●		
United Arab Emirates			●
United States	●	●	●
Total	**34**	**40**	**45**

Table 2.10 Benchmarking participants in PIRLS (2001 – 2011)

	Assessment Year		
	2001	**2006**	**2011**
Abu Dhabi-UAE			●
Alberta-CAN		●	●
Andalusia-ESP			●
British Columbia-CAN		●	
Dubai-UAE			●
Maltese-MLT			●
Nova Scotia-CAN		●	
Ontario-CAN	●	●	●
Quebec-CAN	●	●	●
Florida-USA			●
Total	**2**	**5**	**8**

Table 2.11 Off-grade participants in PIRLS (2001 – 2011)

Off-grade participants	**Assessment year**		
	2001	**2006**	**2011**
Botswana[5]			●
Eng/Afr(5)-RSA[6]			●
Honduras[5]			●
Iceland[1]		●	
Kuwait[5]			●
Morocco[3]			●
Norway[1]		●	
Sweden[4]	●		
Total	**1**	**2**	**5**

● = Indicates participation in particular assessment with results reported.
[1] Administered the PIRLS 4th-grade assessment to 4th-grade students and 5th-grade students in 2006.
[2] Participated in 2001 and/or 2006 but data not comparable for measuring trends to 2011, primarily due to countries improving translations or increasing population coverage.
[3] Administered the PIRLS 4th-grade assessment to a national sample of 4th-grade students and a national sample of 6th-grade students in 2011.
[4] Administered the PIRLS 4th-grade assessment to 3rd-grade students and 4th-grade students in 2001.
[5] Administered the PIRLS 4th-grade assessment to 6th-grade students in 2011.
[6] Republic of South Africa (RSA) tested 5th-grade students receiving instruction in English (ENG) or Afrikaans (AFR).
NOTE: OECD member countries are bolded. Subnational education systems are italicized.

SOURCE: Table adapted from https://nces.ed.gov/surveys/pirls/countries.asp with data obtained by the International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS).

PIRLS Literacy

Countries whose students of the target age are still developing fundamental reading skills can participate in a less demanding test, PIRLS Literacy, which can be administered to students in Grades 4, 5 or 6. PIRLS Literacy stems from prePIRLS which was introduced in 2011. For the 2016 cycle of PIRLS this has been repackaged as PIRLS Literacy. PIRLS Literacy reflects the same conception of reading as PIRLS but it is less difficult and is designed to assess basic reading skills that are a

prerequisite for PIRLS. The reading passages in it are shorter, with easier vocabulary and syntax. Students' strengths and weaknesses in reading comprehension can be measured through their ability to read and comprehend these passages. The IEA introduced PrePIRLS to offer an opportunity for countries with relatively low levels of learning to measure and improve children's learning outcomes systematically (Mullis et al., 2009). The purpose of PIRLS Literacy is to provide better measurement at the lower end of the scale. Countries whose fourth grade students are still developing fundamental reading skills can participate in PIRLS Literacy and obtain results on the PIRLS achievement scale since there are some reading passages and questions common across PIRLS and PIRLS Literacy which enable the two assessments to be linked. Depending on a country's stage of education development and the student's reading level, countries can choose to participate in either or both PIRLS and PIRLS Literacy (Martin et al., 2015).

ePIRLS

In 2016 a third PIRLS assessment, ePIRLS, will be administered for the first time. It has been designed to assess competency in online reading, allowing countries to assess how successful they are in preparing fourth grade students to read, comprehend, and interpret online information. ePIRLS uses a simulated internet environment with school-like assignments about science and social studies topics to measure achievement in reading for informational purposes. ePIRLS is considered an extension of PIRLS and all students participating in this assessment are expected to have participated in PIRLS.

The complete ePIRLS consists of four school based online reading tasks. Each task with accompanying questions takes 40 minutes to complete, same as PIRLS and PIRLS Literacy. Each participating student completes only two ePIRLS tasks followed by a short online questionnaire that takes approximately 5 minutes. This keeps the burden on students to a reasonable limit. There are 12 possible task combinations in ePIRLS and the tasks are distributed randomly to groups of students who are approximately equivalent in terms of ability. Item response theory (IRT) scaling method is used to draw a comprehensive picture of the online informational reading achievement of a country's fourth grade student population. This is made possible by pooling together individual students' responses to the tasks they were assigned. As 2016 is the inaugural year for ePIRLS, not much detail is available on this assessment.

2.6.2 Design and sample

Sample

The international PIRLS target population consists of students enrolled in the grade that represents four years of schooling, provided that the mean age at the time of testing is at least 9.5 years. To better match the assessment to the achievement level of students, countries have the option of administering PIRLS or PIRLS Literacy at the fifth or sixth grade.

Focusing on keeping the burden on schools, teachers and students to a minimum, PIRLS employs thorough school and classroom sampling techniques to measure achievement of the student population accurately by assessing just a sample of students from a sample of schools. PIRLS

employs a two-stage sampling design where a sample of schools is drawn in the first stage and one or more intact classes of students are selected from each of the selected schools in the second stage. As PIRLS follows the same sampling techniques as TIMSS (Joncas & Foy, 2011), details of sampling can be found in Section 2.2.2. For further details on sampling please see 'Sample Design in TIMSS and PIRLS' available online at http://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf.

Test Design

Keeping in mind the broad coverage and reporting goals of PIRLS framework the PIRLS Reading Development group found that a valid and reliable measure would require students to answer questions based on reading passages with a total testing time of eight hours. Considering scheduling and concentration issues, the testing time is limited to 80 minutes per student with an additional 15-30 minutes for a student questionnaire (Martin et al., 2015). To collect adequate data within this short time, PIRLS assessment design uses a matrix sampling technique similar to that used in TIMSS and TIMSS Advanced (discussed in Section 2.2.2 and Section 2.4.2).

For both PIRLS and PIRLS Literacy, each reading passage and its accompanying items is assigned to a block. Each item block requires approximately 40 minutes of student testing time. The blocks are used to construct individual student booklets. Both assessments have 12 item blocks (separate sets for PIRLS and PIRLS Literacy). The 12 item blocks are spread across 16 booklets, each of which contains two item blocks. selected combinations of item blocks are used in the booklets to ensure linking across booklets within PIRLS and to maintain links between PIRLS and PIRLS Literacy. The 16 booklets are distributed among students in participating classrooms to make sure the groups of students completing each booklet are approximately equivalent in terms of ability (Martin et al., 2015).

PIRLS and PIRLS Literacy assessments use two question formats: multiple-choice and constructed-response. The multiple-choice questions are worth one score point each whereas the constructed-response questions can be worth one, two or three score points depending on the depth of understanding required. The selection of format is based on the process being assessed. There is a slightly higher percentage of constructed response items in PIRLS Literacy assessment, comprising up to 60 per cent of the total score points.

In addition to measuring reading ability of children in fourth grade, PIRLS focuses on the home, community, school, and student factors associated with their reading literacy. To fulfil this important purpose, data on the contexts of learning to read are collected through questionnaires completed by students, their parents, teachers, and principals. Moreover, information on national and community contexts for learning are provided by the National Research Coordinators through the curriculum questionnaire and their country's entry in the PIRLS 2016 Encyclopaedia.

2.6.3 Use of data

According to Martin et al. (2015) countries use PIRLS data for a number of purposes including:

i) system-level monitoring of educational achievement in a global context

ii) initiating education reforms when PIRLS achievement results are lower than other countries or lower than expected
iii) making special effort to reduce achievement disparity among social, ethnic, or regional groups
iv) using the data, framework, released items and scoring guides as a basis for updating curriculum and textbook, improving classroom instructions.

2.6.4 Implementation considerations

The PIRLS 2016 projects are funded by IEA and fees from participating countries, with support from the United States Department of Education through the National Center for Education Statistics.

Participation fees are assessed in two currencies and on a yearly basis for each of the five years of the project (2013–2017). The participation fee for PIRLS 2016 is USD 20,000 (EUR 20,000) per year. For PIRLS Literacy, the yearly participation fee is USD 20,000 (EUR 20,000). There is a reduction in fees for countries participating in PIRLS 2016 and PIRLS Literacy together. The fee for ePIRLS is USD 12,500 (EUR 12,500) per year, in addition to the PIRLS 2016 fee.

Activities for PIRLS 2016 started with the national research coordinators' first meeting in February 2013. Framework and instrument development were carried out in 2013-14 and field tests conducted in early 2015. The data collection for the main survey is scheduled to take place in October-December 2015 (southern hemisphere countries) and March-June 2016 (northern hemisphere countries). The next cycle of PIRLS is planned for 2021. If a similar schedule is followed the first meeting of national research coordinators will be in February 2019. Consequently a decision regarding participation would be needed by the end of 2018. Further details on participating in PIRLS can be found online at http://timss.bc.edu/pirls2016/participate.html.

3 Multi-country studies, tools and programs

3.1 South East Asia Primary Learning Metric (SEA-PLM)

The Southeast Asia Primary Learning Metric (SEA-PLM) is an initiative run by the Southeast Asian Ministers of Education Organisation (SEAMEO). The Philippines were part of Phase I of this study, which was designed to measure the learning outcomes of primary school children (starting with 10 year olds). The key domains of the assessment are reading, writing, mathematics and global citizenship/civics education.

The SEAMEO council consists of 11 Southeast Asian education systems: Brunei Darussalam, Cambodia, Indonesia, Lao PDR, Malaysia, Republic of the Union of Myanmar, Philippines, Singapore, Thailand, Timor-Leste, and Socialist Republic of Vietnam. The main goal of SEA-PLM is '*improving quality of education through system level monitoring of learner achievements*' (SEAMEO, 2013).

The SEA-PLM is developing a common tool translated into national languages for each domain, in order to differentiate between lower and higher performing students within countries and also to facilitate exploration of cross-national variations in the South East Asia regional context (UNICEF, 2014).

The SEA-PLM initiative will support SEAMEO member countries to:

"1. better measure and understand the status of learning achievement amongst the general population and for specific groups (e.g., boys/girls; sub-nationally; public/private sectors) through the lens of equity;
2. use culturally appropriate metrics for formative and summative purposes that can assess 21st century skills and critical thinking;
3. heighten the quality of education by making recommendations on areas for improving the relevance and suitability of curriculums in primary school;
4. assert equitable learning environments that correspond with the quality of education and holistic learning approaches as defined by the metric;
5. build technical and analytical capacities of national examination and assessment staff; and
6. strengthen ASEAN technical collaboration on learning assessment and standards across education systems." (UNICEF/EAPRO, Bangkok, 2013).

Currently the SEA-PLM initiative is implementing Phase II; the development of tools and protocols, as well as preliminary testing of the tools. The countries participating in Phase II are Indonesia, Viet Nam and Timor Leste.

The Australian Council for Educational Research (ACER), with support from the UNICEF East Asia and Pacific Regional Office, is working with South East Asian countries through SEAMEO to provide the assessment tools. Following field trials, translations and further refinement, ACER announced that the proposed testing for SEA-PLM will commence in 2016.

3.2 The Early Grade Reading Assessment (EGRA) tool

The early grade reading assessment (EGRA) is designed to test orally the most basic foundation skills for literacy acquisition in early grades (targets grades 1 to 3), including pre-reading skills such as listening comprehension. Note that EGRA is a tool, rather than an assessment program. The test requires about 15 minutes per child and includes timed, 1-minute assessments of letter naming, nonsense and familiar words, and paragraph reading. Additional (untimed) segments include comprehension, relationship to print, and dictation. The assessment was designed as an inexpensive and simple diagnostic of individual students' progress in reading. The aim is for ministry personnel to use the results to identify schools with specific needs and to develop instructional approaches to improve students' foundation reading skills (USAID, 2013).

The EGRA was developed in 2006 by the Research Triangle Institute (RTI) international through funding from the United States Agency for International Development (USAID) and the World Bank (Gove & Wetterberg, 2011). The assessment has been piloted, adapted for use and implemented in more than 60 countries, in 100 languages, as of March 2014, including in the Philippines.

The student instrument includes the oral questions and also a brief questionnaire for the student, to gather basic information about the home and academic environment. In addition to the student instrument, a teacher survey and classroom observation tool are used to help determine factors affecting early literacy achievement.

The data collection for EGRA (in the Philippines) involves the Ilocos Region (Region I), Western Visayas (Region VI), Central Visaya (Region VII) and Maguindaao (ARMM). In 2014, the tests were conducted in these languages: Maguindanaoan (ARMM), Ilokano (Region I), Hiligaynon (Region VI), Cebuano or Sinugbuanong Binisaya (Region VII) (USAID, 2014). Tests are conducted at the end of the school year.

The Philippines administered EGRA in 2009, 2010, 2012-2013, 2013, and 2014:

- 2009: 1426 students in Grade 1 and 3 from 39 schools were sampled. Assessments were conducted in five languages – Tagalog, English, Magindanoan, Ilongo, and T'boli. The purpose of participation in EGRA was to (i) conduct baseline evaluation of learning outcomes of sponsorship-funded programmes in Metro Manila and South Central Mindanao program sites, and (ii) inform Literacy Boost programming in selected communities (ACER, 2013).
- 2010: 780 students from 13 schools were sampled. Participation in EGRA was used (i) as baseline measure for a Whole School Reading Program, (ii) to inform training for teachers, and (iii) to inform instruction (ACER, 2013).
- 2012-2013: 810 students were sampled for pre- and post-testing. For this study, EGRA was a "tool used for a research study on program efficacy" (ACER, 2013).
- 2013: 2810 students were sampled. Assessments were administered in three languages: English, Filipino (Tagalog), and Ilokano.
- 2014: 3200 Grade 1 and 2 students in Ilokano, Hiligaynon, Cebuano and Maguindanaoan were assessed.

3.3 Early Grade Mathematics Assessment (EGMA)

The Early Grade Mathematics Assessment (EGMA) was developed by the Research Triangle Institute (RTI) international, through funding from the United States Agency for International Development (USAID).

The EGMA had been used in 14 countries as of March 2014; the Democratic Republic of Congo, Dominican Republic, Ghana, Iraq, Jordan, Kenya, Liberia, Malawi, Mali, Morocco, Nicaragua, Nigeria, Rwanda, and Zambia. In July 2014 creation of tools for the EGMA for the Philippines was begun.

The core EGMA consists of eight sub-tests; Number Identification, Number Discrimination (reasoning about magnitude), Missing Number (recognition of number patterns), Addition Level 1, Addition Level 2, Subtraction Level 1, Subtraction Level 2, and Word Problems (RTI, 2014).

3.4 Literacy Boost

Literacy Boost is a literacy program designed and implemented by Save the Children. Literacy Boost was launched in Malawi and includes a total of 24 countries around the world (Save the Children, 2015). The goal of this programme is to support the development of reading skills in young children through three steps:

a) Reading Assessments
   Children's baseline and end-line reading levels are measured, along with evaluation of children's literacy and learning needs. Assistance is also given to schools and ministries of education to help track student progress
b) Teacher Training
   Specific training related to incorporating the five core reading skills into students' regular curricula is provided to teachers
c) Community Action
   Parents and communities are given support and resources to help children learn through fun out-of-school literacy activities and locally relevant reading materials.

In 2009, the Philippines participated in the Literacy Boost programme to evaluate programs for disadvantaged students and schools, and to conduct baseline evaluation of early years reading. A total of 1426 grade 1 and 3 students from 31 school participated in this programme.

3.4.1 Literacy Boost Partnership Program

Adapted from Save the Children's Literacy Boost Program, the Literacy Boost Partnership Program is an initiative by World Vision and Save the Children. This three-year program began in October 2011 and was implemented in a selection of World Vision area development programmes (ADPs) in Burundi, Ethiopia, Kenya, Malawi and Rwanda (World Vision International, 2015).

3.5 East Asia Learning Achievement Study (EALAS)

The East Asia Learning Achievement Study (EALAS) was a multi-country capacity building project, funded by UNICEF, designed to assess the learning achievements of primary school-aged children. Studies were undertaken from 2004 to 2006, and conducted at Grades 3 and 5, or Grades 4 and 6. Students were assessed in national curricular domains including mathematics, language, science and life skills. Information about students and schools were also collected. Unlike other multi-country assessments, EALAS does not compare countries as there are no common items in examination across countries. Instead, examinations are based on national curriculum in the testing country. The main focus of EALAS was to bring "greater clarity to the process of measuring learning achievement" (UNICEF, 2007). Assessments were created based on the SOLO (Structure of Observed Learning Outcomes) taxonomy, and analysis was performed using Rasch modelling. Some of the project activities include developing measuring instruments and developing in-country skills in testing and analysis processes.

The Philippines was involved in the EALAS project in 2005. Three domains – mathematics, language and science – were assessed with Grades 4 and 6 students.

3.5.1 Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ)

The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) is a consortium of education ministries who conduct large-scale national research studies to assess students' literacy and numeracy performance levels in Southern and Eastern Africa. Currently, SACMEQ consists of 16 participating Ministries of Education.

The main purposes of SACMEQ are to:

a) "Provide training opportunities that will build the technical capacity of the SACMEQ Ministries of Education to monitor and evaluate the conditions of schooling and the quality of their own education systems.
b) Undertake co-operative educational policy research in order to generate information that can be used by decision-makers to plan the quality of education.
c) Utilize innovative information dissemination approaches and a range of policy- dialogue activities in order to ensure that SACMEQ research results are widely discussed, debated, and understood by all stakeholders and senior decision-makers and then used as the basis for policy and practice" (SACMEQ, 2015).

To date, SACMEQ has successfully completed three large-scale cross-national educational policy research projects. SACMEQ is currently implementing the fourth project. Participation in the SACMEQ research studies is limited to countries in the African region.

3.5.2 Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC)

The Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC), or the "Program on the Analysis of Education Systems", was established in 1991 and has since carried out evaluations in 15 francophone sub-Saharan African countries (Bernard, n.d.).

PASEC is designed to assess student's abilities in mathematics and reading French (Education Policy and Data Center, 2012). These evaluations and assessments are typically conducted on 2nd and 5th grade students at the beginning and end of each school year so that students' growth can be measured over the course of that year.

3.5.3 Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) - Latin American Laboratory for Assessment of the Quality of Education

The Latin American Laboratory for Assessment of the Quality of Education (LLECE) is the network of quality assessment systems for education in Latin America. LLECE is a regional assessment managed by UNESCO's Regional Bureau for Education in Latin America and the Caribbean in Santiago Chile. The assessment began administration in 1997, where students in Grades 3 and 4 were tested. The current research design is students in Grade 3 and Grade 6 every five years. The objectives of the assessments are:

- "producing information about students' learning achievements and analyzing associated-factors that explain this progress;
- supporting and advising the measurement and assessment Units of the different countries; and
- serving as a forum for reflection, debate and exchange of new approaches and focuses on education evaluation." (UNESCO, 2013)

3.6 Household surveys

3.6.1 Multiple Indicator Cluster Survey (MICS)

UNICEF developed the Multiple Indicator Cluster Survey in response to the World Summit for Children to measure progress towards an internationally agreed set of mid-decade goals. The survey was designed to monitor the situation of women and children through an international household survey. The first round of the MICS was conducted in 1995 in more than 60 countries (UNICEF, 2014). By 2015, more than 280 surveys will have been implemented in more than 100 low and middle income countries. Data can be disaggregated by various geographical, social and demographic characteristics allowing UNICEF, along with its global partners, to address the divides and disparities that persist among regions and within countries. Data are collected through face to face interviews in national or sub-national representative samples of households (UNICEF, 2014).

The Philippines participated in the MICS in 1996 (MICS1) and then 1999 (MICS2). The survey for 2015/2016 is MICS5.

3.6.2 Uwezo

Uwezo (meaning 'competency' in Kiswahili) was a five year initiative (2009-2014) that aimed "to improve competencies in literacy and numeracy among children aged 6-16 years in Kenya, Tanzania and Uganda" (Uwezo, 2014). Every year, literacy and numeracy levels of children were assessed using country-wide household based surveys in English, and in the local language. Tests for each year for each country are available from the Uwezo website, www.uwezo.net/. Uwezo operates across East Africa, housed by Twaweza (means 'we can make it happen' in Swahili), a ten year citizen centred initiative, focused on large-scale improvement.

Some of the findings of the study to date are (Uwezo, 2014):

- There were large differences in learner achievement among the three countries, with Kenya performing better than Tanzania or Uganda.
- There were major differences in pass rates among districts within the individual countries, for example Westland – Nairobi had a 87.7% pass rate compared to a 7.2% pass rate for East Pokot – Rift Valley (in Kenya).
- Children from poorer households consistently performed at lower levels on all tests, across all ages.
- The basic competency levels for literacy and numeracy across east Africa have not changed since 2009/2010 (mean test scores remaining constant).
- Teacher attendance rate varies from 82% (Tanzania) to 89% (Kenya and Uganda).

In terms of using the data that was generated from the surveys and tests, the 'Theory of Change' describes the four major stages of the initiative:

1. Annual assessments of country-wide learning;
2. Communicate findings widely and foster broad public debate;
3. Shift from schooling inputs to learning outcomes;
4. Learn, monitor and evaluate.

In a nutshell, the data produced is not explicitly used to inform curriculum reform or policy changes to improve the students learning outcomes; rather, it is anticipated that community groups will encourage a change in the education system, due to the dissemination of poor literacy and numeracy results in these countries.

3.6.3 Annual Status of Education Report (ASER)

The Annual Status of Education Report (ASER) is an annual survey that generates and provides estimates of enrolment and basic learning levels for all children aged 5-16 years in rural India (ASER Centre, 2015). ASER is a household-based survey that has been conducted every year since 2005. Information such as basic household information, parental education and children's schooling status is collected using this survey. In addition, children in the 5-16 years age group are tested in basic reading and basic arithmetic (ASER Centre, 2015).

3.6.4 Literacy Assessment and Monitoring Programme (LAMP)

The UNESCO Institute of Statistics (UIS) developed the Literacy Assessment and Monitoring Programme (LAMP) to provide diagnostic information to monitor and improve literacy skills (UIS, 2014). The surveys are administered through the Ministry of Education to adults in participating countries. The LAMP programme tests literacy in three major domains; continuous text (prose), non-continuous text (documents) and numeracy. Results are reported as a continuum of achievement which is designed to be meaningful to the respondents. The data that is generated from the tests are intended for national and cross-national comparisons.

The three main objectives of LAMP are:

a) Develop a methodology for assessing literacy in developing countries;

b) Provide literacy data to inform the participating countries' policy-making and literacy programme design, and to help international monitoring and policy making;
c) Build statistical capacity in the areas of surveys and of literacy assessment.

As of 2011/12, the following countries had been involved in data collection: Mongolia, Jordan, Palestine, Paraguay, Vietnam, Niger, El Salvador, Morocco, Namibia, Afghanistan, Jamaica, Lao PDR, Nigeria, and India (LAMP Update, 2011). The most recent published update from LAMP was released in 2011. As of July 10, 2015, results were also available for Mongolia, Jordan, Palestine and Paraguay on the UIS, LAMP website; no other country results appear to be available. According to the 2009 LAMP '*Next generation of literacy statistics, technical paper 1'*, the UIS provided the conceptual methodological and technical foundations for LAMP implementation (occurring between the years 2003 and 2008), but the national implementation of LAMP "falls beyond what the UIS can afford" (UIS, 2009). Countries were expected to fund their national involvement.

Kenya reported that the LAMP developmental phase was "too long" and that "a critical analysis of 12 filter module assessment questions…revealed that 90% of these items would not be culturally relevant to the Kenyan situation" (Kebathi, J., 2008). Kenya went onto develop their own adult literacy assessment with the methodological skills they developed as part of their involvement in both LAMP and SACMEQ (Southern and Eastern Africa Consortium for Monitoring Educational Quality), developing assessment items that were more specifically tailored to the Kenyan culture and context (Kebathi, J., 2008).

The country summaries produced by the programme contain a broad synthesis of the data collected, with three levels of competency reported for each of the major domains (continuous text (prose), non-continuous text (documents) and numeracy). The proportion of the survey population on each level of competency is reported in terms of gender, age and education level.

3.7 Other

3.7.1 Learning Metrics Task Force (LMTF)

The Learning Metrics Task Force (LMTF) was convened in 2012 to investigate how learning progress can be tracked at a global level and to "improve the learning outcomes of all children and youth by strengthening assessment systems and the use of assessment data" (UNESCO Institute for Statistics, 2014). LMTF is run by the UNESCO Institute for Statistics (UIS) and Center for Universal Education (CUE) at the Brookings Institution.

To date, LMTF has conducted two phases of research work. In the first phase (LMTF 1.0), the task force completed several rounds of global consultation and technical development involving 1,700 people from 186 countries. This consultative process was structured and guided by three research questions:

a) What learning is important globally?
b) How should it be measured?
c) How can measurement of learning improve education quality?

Through this consultation process, a series of recommendations for improving learning outcomes and measurement at the global level was put forward. Since July 2014, the task force has engaged in a second phase of work (LMTF 2.0) to implement key LMTF recommendations from the first phase. Fifteen countries were selected as "Learning Champions" to participate in LMTF 2.0; national stakeholders will be working to adapt and implement these recommendations to their national contexts. Applications for participation closed in May 2014 and are no longer being accepted.

3.7.2 Russian Education Aid for Development (READ) Trust Fund

The Russian Education Aid for Development (READ) trust fund was created in October 2008, as a partnership between the Government of Russia and the World Bank. The $32 million dollar trust fund was dedicated to helping developing countries improve student outcomes. The aims of the fund were to develop the capacity of low-income countries to assess student learning and to use the information from those assessments to improve teaching practises.

According to the World Bank (World Bank Group, 2015), the READ trust fund was designed to help countries:

- "Establish or strengthen existing systems or institutions that formulate learning goals and carry out assessments of student learning;
- Improve existing or develop new instruments to measure student learning outcomes; and
- Strengthen existing or develop new policies to use learning outcomes data to improve teaching and learning."

Eight countries were selected upfront by Russia for READ assistance. These countries worked with World Bank operational teams to develop country specific programs of READ assistance based on the interests of the country.

The participating countries were: Angola, Armenia, Ethiopia, the Kyrgyz Republic, Mozambique, Tajikistan, Vietnam, and Zambia. The countries utilised the trust fund in different ways. Some chose to focus on a specific issue such as developing a national large-scale assessment program or a establishing a new testing centre, while other countries opted to address gaps in multiple areas.

The knowledge dissemination process for the READ trust fund has included publication of working papers and reports on developing student assessment approaches in various countries. The publications are available from the website http://www.worldbank.org/en/programs/read#4.

4 Comparisons between PISA, TIMSS and PIRLS

This section focuses on differences between three international large-scale assessments – PISA, TIMSS and PIRLS. While all three assessments involve school-aged students, PISA, TIMSS and PIRLS differ in several ways (see Table 4.1).

First, PISA differs from TIMSS and PIRLS in that assessment is age-based; it samples 15-year old students regardless of the number of years students have received formal schooling. In contrast, TIMSS and PIRLS are grade-based assessments where children in specific grades or year levels are sampled and assessed. PISA, TIMSS and PIRLS are administered to a sample of students to allow results to be generalised to the larger population:

- PISA: 15-year-old students
- TIMSS: Grade 4 and 8 students
- TIMSS Advanced: Grade 12 students
- TIMSS Numeracy: Grade 4, 5 or 6 students
- PIRLS: Grade 4 students
- PIRLS Literacy: Grade 4 students

Second, each assessment provides different data and information. PISA is an assessment of 15-year old students' performance in reading, mathematics, and science. PISA also includes assessment and measures of other general competencies such as learning strategies, collaborative problem solving and financial literacy. More specifically, PISA assesses students' ability to apply these skills and knowledge in real life situation and contexts. On the other hand, TIMSS and PIRLS provide data and information on trends and performance of students in specific domains, such as mathematics, science and reading achievement.

Third, the frequency and period of testing for each assessment is different. PISA, TIMSS and PIRLS are conducted every 3, 4 and 5 years, respectively. Additionally, testing is conducted at different times throughout the year for each international large-scale assessment programme.

Last, the cost to participate in each large-scale assessment programme is different. As shown in Table 4.1, the participation fee for PISA is substantially higher than TIMSS or PIRLS. Nevertheless, regardless of assessment programmes, participating countries are responsible for the costs of administration and implementation at the national level (e.g., cost for hiring research personnel and statistician) and also contribute to the costs of coordinating the study internationally.

Table 4.1 Overview of PISA, TIMSS and PIRLS

Assessments	PISA	TIMSS			PIRLS		
		TIMSS	TIMSS Advanced	TIMSS Numeracy	PIRLS	PIRLS Literacy	ePIRLS
Occurrence	Every 3 years	Every 4 years	1995, 2008, 2015	Starting 2015	Every 5 years	Starting in 2016	Starting in 2016
Dates of next study	2015 2018 2021	2015 2019	2015	2015	2016 2021	2016	2016
Who is tested?	15-year-old students	4th and 8th grade students	12th grade students (final year secondary students)	4th, 5th or 6th grade students	4th grade students (can also be administered to 5th and 6th grade students)	4th grade students	4th grade students
What is tested?	Ability to apply skills and knowledge in real life contexts	Mathematics and science	Advance mathematics and physics	Numeracy learning outcomes, including fundamental mathematical knowledge, procedures and problem-solving strategies	Reading achievement	Reading achievement; a less difficult version of PIRLS	Computer-based reading assessments; online reading skills and competencies
How much does it cost*?	EUR 182,000 (payable over four years at EUR 45,500 per year)	USD 25,000 per year per grade^	USD 37,500 per year	Part of TIMSS 2015	USD 20,000 per year+	USD 20,000 per year+	USD 12,500 per year (in addition to the PIRLS fee)

*Prices as of June 2015.

^There will be a reduction in fees for countries participating in TIMSS 2015 for two grades or both TIMSS 2015 and TIMSS Advanced 2015.

+There will be a reduction in fees for countries participating in both PIRLS and PIRLS Literacy

5 Considerations

5.1 Scheduling of assessments

Prior to the K – 12 education reform and through to 2015, Filipino students have been required to complete a large number of assessments in their Grade 1 – 10 education. Not all countries require their students to complete as many assessments. According to the 2013 survey conducted by UNESCO, the Philippines had the second highest number of national examinations after Thailand, as seen in Table 5.1 (UNESCO, 2013).

Table 5.1 Approximate student ages at national examinations in the Asia-Pacific

Country	Age												
	5	6	7	8	9	10	11	12	13	14	15	16	17+
Bhutan												●	
Cook Islands				●	●						●	●	●
Iran							●	●	●	●	●	●	
Kazakhstan											●	●	●
Kyrgyzstan							●	●	●	●	●	●	
Lao PDR						●				●			●
Mongolia*					●	●	●			●	●	●	●
Myanmar			●	●	●	●	●	●	●	●	●		
Nepal			●			●			●		●	●	
New Zealand											●	●	●
Palau		●	●	●	●	●	●	●	●		●		
Philippines		●	●	●	●	●	●	●	●	●	●	●	
Sri Lanka						●	●					●	
Thailand		●	●	●	●	●	●	●	●	●	●	●	●
Tokelau			●	●		●	●				●	●	
Uzbekistan					●	●	●	●	●				●
Victoria (AUS)													●

The Philippine national assessment schedule should be taken into account when looking at international large scale assessment possibilities. The timing of assessments has implications for infrastructure and staffing, as well as for the possible analysis of the large scale assessment results against any national assessment data. Refer to Table 5.2 for key dates of assessment data collection.

Table 5.2 Key Dates (1-12, Kindergarten discounted)

	International					Multi-Country			National Assessments			
Assessment	**PISA**	**TIMSS**	**TIMSS-Advanced**	**TIMSS-Numeracy**	**PIRLS**	**SEA-PLM**	**EGRA**	**EGMA**	**NAT**	**LAPG**	**PIRI**	**NCAE**
Occurrence	**3 years**	**4 years**	**Sporadic**	**4 years**	**5 years**	**Study**	**Study**	**Study**	**Yearly**	**Yearly**	**Yearly**	**Yearly**
Dates of next study	**2015** **2018** **2021**	**2015** **2019**	**2015**	**2015** **2019**	**2016** **2021**							
Content area/s	Reading, Mathematics, Science, Collaborative Problem Solving, Financial literacy	Number, algebra, and geometry in mathematics, and earth science, biology, and chemistry in science	mathematics and physics	fundamental mathematical knowledge, procedures, and problem solving strategies	PIRLS literacy – Fundamental reading skills PIRLS – Reading skills	Reading, writing, mathematics global citizenship/ civics	Early reading	Early Mathematics	National achievement test	Reading proficiency 19 MT languages	Reading proficiency English/ Filipino	National Career Assessment Exam
Target Age	15 year olds (Grade 10)	Grades 4 and 8	Grade 12	Grade 4, 5 or 6	Grade 4, 5 or 6	10 year olds (Grade 4)	Grades 1 to 3	Grades 1 to 3	Grades 3, 6 and 10	Grade 3 (public schools)	Grades 1 to 6	Grade 9
Testing Date*	March to August	March–June (northern hemisphere) October–December (southern hemisphere)	March–June (northern hemisphere) October–December (southern hemisphere)	March–June (northern hemisphere) October–December (southern hemisphere)	March–June (northern hemisphere) October–December (southern hemisphere)	Start 2016 Dates unknown	March each year	March 2016?	March each year	March each year	January each year	August each year
Sign up Date*	End of 2014 (2018) **End 2017** (2021)	End of 2013 (2015) **End of 2016** (2019)	Not released	**End of 2016** (2019)	2016	N/A	N/A	N/A	N/A	N/A	N/A	N/A

PISA - Programme for International Student Assessment
PIRLS - Progress in International Reading Literacy Study
EGRA - Early Grade Reading Assessment
NAT - National achievement test
PIRI - The Philippine Informal Reading Inventory
* Approximate based on previous years

TIMSS - Trends in International Mathematics and Science Study
SEA-PLM - South East Asia Primary Learning Metric
EGMA - Early Grade Mathematics Assessment
LAPG - Language assessment for Primary Grades
NCAE - National Career Assessment Exam

5.2 Stage of curriculum implementation

In evaluating the results from ILSA, the context which pertained to the students' educational experiences needs to be considered. Countries may decide to engage in ILSA for a variety of reasons. For example, does a country embark on ILSA to obtain insight into the status quo with a view to implementation of reform? Or does the country wish to review progress or changes in educational performance that might have ensued from education reform? Or does a country wish to review progress of students against previously established benchmarks in the country? In the case of the Philippines, these questions are particularly salient.

The country is some two years into K – 12 implementation. Is the country looking for affirmation of the need for the reforms, for the effectiveness of the reforms, or for benchmarking data for future use? Obviously these options are not mutually exclusive, but it is important that the questions of interest are identified a priori, with discussions of the implications of results for each question, as explored in brief in Table 5.3.

Table 5.3 Possible arguments

	Possible claims if ranking is higher than designated benchmark countries	Possible claims if ranking is lower than designated benchmark countries
Purpose		
Affirmation of need for reform	Reform is not justified since the cohorts being assessed have not experienced the full K – 12 and country results are already fine	Reform is justified
Affirmation of effectiveness of reform	Reform is justified – without reference to the fact that cohorts assessed have not experienced the full K - 12	Reform is not yet fully implemented, so this is an expected result
Use for benchmarking within country	Greater pressure for higher results next time to demonstrate effectiveness of the reform	Demonstrates starting point in the reform effort, with assumption of improvement next time around

The status of the K to 12 curriculum reform implementation should be considered in regards to the next dates of large-scale assessment data collection. The following sections refer to the stage of curriculum that students will have completed when the next possible assessment rounds take place. Student achievement at any one point in time is an indicator of a cumulative effect of the educational experience of that student. It is important that the educational context which each student has experienced is considered when interpreting large scale assessment results.

5.2.1 PISA

PISA 2018 data collection (the next available study for involvement) for 15 year old students would likely involve students who began Grade 1 in the SY 2008/2009 - these students miss the

implementation of the elementary part of the K to 12 curriculum by four years. This cohort of students however, began Grade 7 in the SY 2014/2015, and will therefore have received the K - 12 curriculum for Grades 7 to 10. The results achieved if the Philippines participate in the 2018 PISA data collection will reflect a mixed education background, with consequent difficulties for attribution of progress to the previous or current education system. High scores could be attributed to the success of the new curriculum, or a reflection of having a solid basis built by the previous curriculum. Low scores could be attributed to students not having been adequately prepared for the new curriculum due to their elementary school experience, or to a failure of the reform. In short, the 2018 PISA data collection occurs during the transition phase between the old and new education systems, with students completing half of each. This timing therefore offers both opportunities and challenges. Accordingly, it is important that the country considers the possible outcomes of the assessment, and proposes a number of hypotheses to investigate in order to be prepared to deal constructively with the direction of the results.

5.2.2 TIMSS

The next TIMSS and TIMSS numeracy assessments' begin data collection in 2019. TIMSS and TIMSS numeracy can be used to test students in Grade 4, who will have begun Grade 1 in the SY 2015/16; these students will be completing the K - 12 curriculum, which will have had four years for teachers and governing systems to establish implementation. Thus a TIMSS or TIMSS Numeracy collection for Grade 4 students in 2019 would provide an accurate representation of the state of students' achievement in mathematics after the implementation of the K - 12 curriculum. TIMSS numeracy assessments are also available for Grade 5 and 6, which is the cohort of students beginning in SY 2014/2015 and 2013/2014 respectively. Each of these cohorts is potentially disadvantaged due to receiving less of the K - 12 curriculum before testing, but advantaged due to being older while completing the assessment (TIMSS numeracy is an easier version of TIMSS, outlined in Section 2.3).

TIMSS is also available for students in Grade 8. The cohort which would be tested in 2019, will have begun Grade 1 in the SY 2011/2012, and will have completed the same part of the curriculum as the cohort of students that could potentially participate in the PISA in 2021. This is the cohort that will receive the secondary K - 12 curriculum, but not the elementary. Thus these students have only completed Grades 7 and 8 of the new curriculum. The consequences of testing this cohort are similar to being involved in PISA – interpretation of the data would not be straightforward due to the multiple curricula experienced by the students. The usefulness of the results could be questionable if it is not possible to explain these as influenced differentially by one part of the educational experience or another. Politically the implications of receiving such results should be considered.

TIMSS Advanced assessment data collection dates for later than 2015 have not yet been released, but if the Philippines are considering involvement, stage of curriculum should also be considered for the next cohort of students to be tested.

5.2.3 PIRLS

The Philippines could sign up for the 2021 round of PIRLS (and possibly in the same year ePIRLS and/or pre PIRLS literacy depending on whether the IEA follows the same assessment schedule). The testing cycle would involve the cohort of student completing Grades 4, 5 or 6 in the SY 2020/2021, which began Grade 1 in SY 2017/2018, 2016/2017, or 2015/2016 respectively. These students will

have received the full K - 12 curriculum, with the full curriculum having been implemented for 3 years at that stage (Grade 6 cohort). Thus students' scores for the 2021 PIRLS could be attributed to the K - 12 curriculum without influence from the previous curriculum. PIRLS would then run five years after, so a 2021 assessment round could provide a baseline if further changes to the curriculum are recommended in the future.

5.2.4 Summary

The Philippines could consider involvement in the Grade 4 TIMSS data collection and/or the 2021 Grades 4, 5 or 6 PIRLS collection, with a view that student results could be interpreted as resulting from the new curriculum (Table 5.4). The PISA, Grade 8 TIMSS, and TIMSS Advanced data would pose interpretation difficulties as data would be collected from a cohort of students that had studied under both the previous curriculum and the current. How the results of assessments could be used or interpreted is an important consideration when deciding on involvement with studies that are designed primarily to benchmark countries against one another.

Table 5.4 Cohorts by year within the context of ILSA

PISA			TIMSS						PIRLS				
2008	2009	1							2011	2012			
2009	2010	2							2012	2013			
2010	2011	3	2011	2012				1	2013	2014			
2011	2012	4	2012	2013				2	2014	2015			
2012	2013	5	2013	2014			1	3	2015	2016			1
2013	2014	6	2014	2015		1	2	4	2016	2017		1	2
2014	2015	7	2015	2016	1	2	3	5	2017	2018	1	2	3
2015	2016	8	2016	2017	2	3	4	6	2018	2019	2	3	4
2016	2017	9	2017	2018	3	4	5	7	2019	2020	3	4	5
2017	**2018**	**10**	**2018**	**2019**	**4**	**5**	**6**	**8**	**2020**	**2021**	**4**	**5**	**6**

5.3 Previous Philippines large-scale assessment

Previously the Philippines has been involved in TIMSS, TIMSS Advanced and some multi-country or regional assessments. The involvement with TIMSS included pilot work in 1993 as a Summer Program where items were administered to teacher candidates, then TIMSS 1999 (Score 345), 2003 (378), and TIMSS Advanced 2008 (355). In the latter, the Philippines participated in the mathematics strand only with a total of 4091 students in their final year, from 118 science high schools, tested. Despite the student population being the higher achievers of the country from the top schools, the Philippine students scored lower than the other participating countries. TIMSS results were well below average.

The agencies involved in the implementation were the Department of Science and Technology (DOST), the Department of Education, and the National Institute of Science and Mathematics Education (NISMED). From the early participation in TIMSS through to 2003, approximately 50% of the same personnel remained involved, down to about 30% by 2008. In terms of technical responsibility, the sampling of schools was undertaken by IEA Central (in Boston), with NISMED providing the schools and regions database for the purpose. The Philippines was responsible for

provision of school and student data, the actual assessment administration, then encoding (managed by the Computer Science group) and scoring only. The Test Administrators were provided by University of the Philippines International School (UPIS), NISMED, the University of the Philippines College of Education, and DOST.

5.3.1 Use of the results

System level

Apart from the international and national report generation, NISMED staff also completed item analyses by regions. The results of these were presented to ManCom meetings in order to contextualise the student performance by describing the sequence of topics taught to students at various times in their curricular experience.

At different stages, TIMSS-like assessments were developed in the Philippines with up to 3 of 17 regional directors adopting new approaches to assessment.

Presentations at DepEd led to some changes in assessment policy; and the generation of DepEd memos to all regions about assessment practices.

School level

In 2003, about N = 5000 booklets were distributed to schools with indicative items so that schools could benefit from what had been learnt through the TIMMS experience. Another booklet was developed which provided strategies on how to teach topics on which the Philippine students had performed particularly poorly. These publications were used by teams which went out to schools to present on the results as professional development activities.

There is no doubt that feedback to schools and teachers about student performance on TIMSS provided valuable information about the reasons for poor performance. The topics upon which test items were based had been included in the curricular studies, so alignment of content was not at issue. The crux of the poor performance was perceived by science and mathematics experts to be student lack of understanding of the concepts that underlie the curricular topics. It appeared that the topics were taught in a rote fashion which did not enhance student understanding. Therefore when items were presented to students in formats and styles that were unfamiliar to them, they did not have the capacity to apply their topic based learning. This issue goes to the content focus of the previous curriculum as opposed to the greater focus on understanding and application characteristic of K – 12. Particularly in maths and science, the view is that the previous curriculum taught content knowledge relevant to college entry rather than for mathematical or scientific literacy.

Sustainability

The training opportunities for teachers that were provided through the analysis of results from TIMSS were of great value. Issues around the sustainability of the gains centre around the retention of teachers and science coordinators at the school level. Many, naturally, are promoted out of these contexts over time. A related issue concerns the manner of training. Where a presenter develops depth of understanding of the links between strengths and weaknesses in student performance, and particular teaching needs and strategies, that understanding tends to reside with the demonstration

teacher. The cascade model of teacher professional development diminishes the learning transfer. Sustainability depends on both with Central DepEd office personnel as well as Regional personnel.

The Philippine TIMSS experiences were useful in a variety of ways. In terms of expectations about the benefits of LSA participation, these may be categorised as providing information to:

- inform policy
- provide capacity building in assessment methods
- guide teachers about the areas of strength and weakness in student outcomes
- alert the country to the realities of education outcomes.

Greater use of the program and therefore greater benefit could eventuate if:

- Engagement from both the educational and political sectors was ensured before starting
- A dedicated unit within DepEd was established on a permanent basis, to implement the LSA, and to act simultaneously as a skills repository for the assessment division/group within DepEd
- A schedule of capacity building and stakeholder engagement activities was provided to support skills development at regional level, and across teacher education colleges and universities.

5.4 Use of assessment data

Participation rates in international large-scale assessments such as PISA and TIMSS, are increasing, particularly in the Asia-Pacific region (UNESCO, 2013). International and multi-country assessment data provide rich information to participating countries, and the usefulness of data is dependent upon the types of data collected, the data analyses that follow, and the dissemination and reporting of such data.

For developing countries such as the Philippines, assessment data have several important uses and purposes. First, assessment data serve as an important database for countries with less well developed education information management systems (UNESCO, 2013). Most ILSA collect data at student, school, regional and national levels. More specifically, these data provide information across factors such as students' performance in specific subjects or a specific learning areas, profiles of high- and low-achieving students, socio-demographic characteristics of students, performance level and profiles of schools involved. Collection of such data could assist in developing and establishing a more efficient educational data management and reporting system. In addition, these assessment data could enable detailed analysis at national, regional and other levels, and provide snapshots of specific groups of students, school and/or overall national performance.

Second, assessment data can be used to review and track national progress and performance. Although some countries such as the Philippines are already systematically collecting national assessment data, participation in ILSA can be a way of analysing the performance of a countries' education system. Participation in ILSA can not only provide representative national data, but also allow for comparison with other countries in terms of performance and ranking in a specific region

or in relation to other participating countries. Given that some ILSA measure performance of a specific sample (e.g., 15-year-olds for PISA), comparison of performance in a particular learning area (e.g., mathematics or science) can be made between groups of students from different schools and even different countries, although careful considerations of other confounding factors must be taken. Furthermore, when a country participates in an ILSA program for more than a year, the results and data can be used to monitor and track cohorts of students, or overall national performance over time. Information obtained from progress monitoring and tracking have valuable implications for policy and programme design and evaluation.

Third, national and international assessment data can be used as powerful and effective tools for reviewing education curriculum or policies, including decisions about professional development and training for teachers, and intervention programs for students.

Various academic articles provide advice on how to accurately use and interpret these data. Rutkowski and others (2010) formed recommendations for researchers wishing to analyse large data sets while avoiding bias. Recommendations included using appropriate sampling weights for the research question, using plausible values with survey software, resampling variance estimations when able, reporting teacher-level data as attributes of the student and not making causal inferences from cross-sectional studies such as TIMSS, PIRLS and PISA (Rutkowski, 2010). DeMars (2015) showed how variance components can be estimated from the sparse data matrices often produced by large-scale assessment data. Chow and Kennedy (2014) used cluster analysis as a secondary data analysis technique, to investigate Asian students' attitudes to their future civic participation, which could not have been achieved using a conventional variable approach. These authors discuss the advantages of secondary data analysis claiming this is an important use of large-scale assessment data. Lewis and Lingard (2015) discuss the multiple effects of ILSA on education policy and research, but like many others, explain how data could be used, rather than how it is being used in practice.

5.4.1 Country use of data

So how are different countries in the Asian region actually using the data? In a 2013 survey of 17 countries (out of 48 countries) of UNESCO's Asia-Pacific Member States, nine countries responded to a question relating to the activities or actions that followed the results of the most influential international large-scale assessments (UNESCO, 2013). The Philippines did not respond to this question in the survey, despite its TIMMS involvement. The remaining seven countries did not respond as they had not participated in any large-scale assessments, rendering the question irrelevant. The activities or actions taken following participation in ILSA are summarised in Table 5.5. These responses indicate that countries vary in their actions and dissemination processes after receiving results from large-scale assessments. The most common actions taken are to review or change the curriculum, or to conduct seminars and conferences for policy makers and researchers. Five countries reported that participation in ILSA led to either a review of or changes to education curriculum. Two to four countries reported that some sort of intervention program was introduced following the results from the most influential international large-scale assessments. The least common action involves feedback to students.

It is not only the student assessment data that can be used to inform policy decisions; most large-scale assessments incorporate both teacher and student questionnaires, which provide a rich data-set for countries to use. Questionnaire data includes variables such as teacher's opinions of their mathematical teaching ability, students' calculator usage, index of home educational resources – the list is almost endless, and different or new questions are added each year. Often the students' scores for particular groups based on these data are compared, so countries can see effects of the variables. For example, one such comparison showed that increased time in class spent on mathematics is not associated with higher scores (TIMSS, 1999). The use of the questionnaire data depends on the country. Each country chooses how to interpret the data produced, and whether to use the information to inform policy changes.

Table 5.5 Actions taken following the results of the most influential international large-scale assessments

Action	Countries reporting action
Review of or changes to the curriculum	Kazakhstan (also change in study plans), Kyrgyzstan, Mongolia, Iran, New Zealand (indirect), Thailand
Intervention programmes for specific group of students	Kyrgyzstan, Myanmar, Thailand
Intervention programmes for specific type or group of schools	Kyrgyzstan, Myanmar
Intervention programmes on specific theme/learning or subject area	Kyrgyzstan, Mongolia, Myanmar, Thailand
Professional development of teachers	Kazakhstan, Kyrgyzstan, Myanmar, Iran, New Zealand (indirect), Thailand (indirect)
Professional development for principals/school leaders	Kazakhstan, Kyrgyzstan, Myanmar
Seminar/conferences for policy-makers and/or researchers	Kazakhstan, Kyrgyzstan, Mongolia, Myanmar, Iran, New Zealand, Thailand
Seminar/conferences for unions and professional bodies	Kazakhstan, Myanmar, Thailand
Feedback to students	Myanmar

In summary, countries use the data gained from these large-scale assessments in different ways, but it is common for countries to presume that policy making changes will occur as a direct result of increasing the 'visibility' of student outcomes between participating schooling systems, as was suggested by Lewis and Lingard (2015).

An external evaluation of the impact of PISA on country policy was reported by OECD (2008). It was reported that no stakeholder group (including the business community, teacher associations, academics and researchers, parents, school principals, local government officials, policymakers) assumed significant responsibility for country outcomes. In fact, 32% of local government officials and 24% of policymakers claimed responsibility, with about 2% of school principals and teacher associations claiming responsibility. It is essential to engage stakeholder groups to realise their influence on student outcomes, and acknowledge their responsibility for these.

5.4.2 Large-scale assessment as a change agent

> ".. PISA .. has had a progressively larger impact on the educational policies of countries. Many countries felt PISA gave them an honest view of where they were in their aspirations to have the best possible talent development. It was not always a happy view. Sometimes it confirmed earlier fears that the country had fallen off track. Sometimes the PISA results were in sharp contrast to previously held beliefs in the quality of a country's education system. This PISA-shock has spurred a rapid change in country policies, with a likely unprecedented upward spiral in the quality of education." (Ritzen, 2013, p. 13)

Ritzen makes the point that iteration, or the cycle of measurement, is a necessary condition for assessment to play an improvement function. Although one-off participation in ILSA programs may well provide benchmark data, its usefulness is severely limited if it is not repeated. In the Philippines case, this is particularly so. Were the Philippines to participate in 2018 PISA for example, this will be at a moment of transition in the education system of the country, with the K – 12 reform underway but not completed. As such, the picture of the educational outcomes for students at that stage will be subject to some speculation concerning impact of the reform. At the same time, collecting benchmark data at that point will provide an excellent opportunity to explore the influence of the reforms in 2021. Accordingly, it is recommended that if the Philippines participate in PISA in 2018, this should be with the intention and explicit plan to participate also in the 2021 round.

5.5 Concluding remarks

The Philippines has recently implemented the Enhanced Basic Education (K - 12) Program, and over the next few years, cohorts of students will be transitioning from the previous curriculum to the newly implemented program, while other students experience only the K - 12 curriculum. During this transition period, students, teachers and schools will no doubt be faced with new challenges.

It is important that the Philippines is clear about the purpose of their participation in ILSA, and that the country considers a set of questions for exploration as part of such participation. Clarity around the purpose needs to widely disseminated in order to engage stakeholders and to pre-empt possible outcomes so that these can be managed constructively. Questions for exploration need to be developed since additional survey questionnaire items may need to be designed in order to ensure that factor level data will be available for use to interrogate patterns in student performance outcomes. Apart from the 'big' questions concerning national performance and its relationship to K – 12 reform, there are questions that can provide input to consideration of the specific features of the K – 12 reform, including impact of the spiral approach on numeracy skills in the event of PISA participation, or impact of an anticipated change of focus from content to understanding and application skills on problem solving, etc. Note that PISA in particular emphasises students' ability to forumlate, solve, and interpret mathematical problems in real life situations. The items are not designed merely for memory recall, and recall alone will not equip students to experience success with these items.

At the same time, the Philippines could usefully explore some of the reasons for varied outcomes in performance in countries where the mother tongue is not the primary language used throughout formal education. In particular, given that English is the language of instruction in the Philippine secondary school classroom, some consideration of the impact of reading proficiency on maths and science scores should be given. There is extensive research relevant to this issue, from the impact for early learners (Robinson, 2010), to impact for performance of older students on PISA assessment (Ercikan et al., 2015). Although ILSA such as PISA follow a rigorous process in finalising items for inclusion at country level, there remains evidence that limited English proficiency has significant implications for students' success in mathematics and science assessments (eg. Noble, Risebery, Suarez, Warren & O'Connor, 2014).

The usefulness and impact of ILSA depends on the degree to which participation provides transparency around the educational achievements of a country's students, and therefore of the education system. Transparency should stimulate reflection and review on the part of policymakers and other stakeholders in education. The essential conditions for such an outcome are outlined by Braun (2013, p. 151-152):

- "The reported outcomes are considered credible, relevant, and sufficiently accurate
- There is acknowledgement of the correspondence between these outcomes and the national goals
- The interpretations of the outcomes, both absolutely and comparatively, are approximately correct
- Stakeholders are inspired by the results, as well as the accompanying public reaction, to propose new policies and allocate resources
- Policymakers maintain a sustained but flexible focus on these policies."

6 References

Introduction References

Benavot, A., & Tanner, E. (2007). The growth of national learning assessments in the world, 1995–2006. In: Education for All by 2015: will we make it? EFA global monitoring report. Paris: UNESCO, 1–17.

Best M, Knight P, Lietz P, Lockwood C, Nugroho D, Tobin M (2013). The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing countries. Final report. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Care, E., & Beswick, B. (2016). Comparative approaches in education. In D. Wyse, L. Hayward, & J. Pandya (Eds.). Handbook of Curriculum, Assessment, and Pedagogy. SAGE Publications.

Greaney, V., & Kellanghan, T. (2008). Assessing national achievement levels in education. Retrieved http://www.uis.unesco.org/Education/Documents/assessing_national_achievement_level_Edu.pdf

Government of the Philippines (2014). The K–12 Basic Education Program. Retrieved from http://www.gov.ph/ K–12/

Republic of the Philippines, (July 2012). An act enhancing the Philippine basic education system by strengthening its curriculum and increasing the number of years for basic education, appropriating funds therefor and for other purposes. Retrieved from http://www.gov.ph/downloads/2013/05may/20130515-RA-10533-BSA.pdf

Ritzen, J. (2013). International large-scale assessments as change agents. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.) The role of international large-scale assessments: perspectives from tehcnology, economy, and educational research. Doredrech: Springer.

PERLS Reading (2015). Programmed English Reading Language System. Retrieved June 17 from http://www.perlsreading.com/overview/

UNESCO (2015). Education For All (EFA) Global Monitoring Report 2015, Achievements and Challenges. Retrieved from https://en.unesco.org/gem-report/#sthash.Us1460EK.dpbs

Wagner, D. A., Lockheed, M., Mullis, I., Martin, M. O., Kanjee, A., Gove, A., & Dowd, A. J. (2012). The debate on learning assessments in developing countries. Compare: A Journal of Comparative and International Education, 42(3), 509-545.

PISA References

Adams, R. J., Wu, M. L., & Carstensen, C. H. (2007). Application of multivariate Rasch models in international large-scale educational assessment. In M. v. Davier. C. H. Carstensen (Ed.), Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 271-280): Springer Verlag.

Adams, R., Berezner, A., Jakubowski, M. (2010) Analysis of PISA 2006 preferred items ranking using the percent-correct method. OECD Education Working Papers, No. 46.

Goldstein, H. (2004) International comparisons of student attainment: some issues arising from the PISA study. Assessment in Education, 11, 319-330.

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007) Translation Equivalence across PISA Countries. Journal of Applied Measurement 8(3) 2007, 249-266.

Kreiner, S & Wuttke, J. (2007), Uncertainties and Bias in PISA: Translated from German article appearing in 'PIZA zufolge PISA' – PISA according to PISA. ISBN 978-3-8258-0946-1. Retrieved from http://www.oxydiane.net/IMG/pdf/Uncertainties_and_Bias_in_PISA.pdf

Le, L (2006b). Investigating Gender Differential Item Functioning Across Countries and Test Languages for PISA Science Items. Paper presented at 5th Conference of International Test commission, Brussels, July 2006.

Le, L.T. (2006a). Analysis of Differential Item Functioning. Paper presented at the annual meeting of American Educational Research Association, San Francisco CA. 8

Le, L.T. (2009). Investigating Gender Differential Item Functioning across Countries and Test Languages for PISA Science items. International Journal of Testing, 9, 2, 122-133.

Le, Luc T. (2007). Effects of item positions on their difficulty and discrimination - A study in PISA Science data across test language and countries. New Trends in Psychometrics. Proceedings of Conference of Psychometric Society, Tokyo, 2007.

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29:133

OECD (2009), PISA 2006 Technical Report, OECD, Paris.

PISA (2015a). PISA FAQ. Retrieved from http://www.oecd.org/pisa/aboutpisa/pisafaq.htm

PISA (2015b), How to join PISA. Retrieved from http://www.oecd.org/pisa/aboutpisa/howtojoinpisa.htm

PISA (2013), Improving learning outcomes worldwide: How PISA can help. PISA for Development. Retrieved from http://www.oecd.org/pisa/aboutpisa/pisa-for-development-brochure.pdf

PISA (2015), PISA for Development national project managers 2015-2018. Retrieved from http://www.oecd.org/pisa/aboutpisa/pisafordevelopmentnationalprojectmanagers2015-2018.htm

TIMSS References

Foy, P., Brossman, B., & Galia, J. (2013). Scaling the TIMSS and PIRLS 2011 Achievement Data. Retrieved from http://timss.bc.edu/methods/pdf/TP11_Scaling_Achievement.pdf

Joncas, M., & Foy, P. (2012). Sample design in TIMSS and PIRLS. Retrieved from http://timssandpirls.bc.edu/methods/t-sample-design.html.

Martin, M. O., Mullis, I. V. S. and Foy, P. (2013) TIMSS 2015 Assessment Design in Mullis, I. V. S., and Martin, M. O. (Eds.), TIMSS Assessment Frameworks. Massachusetts: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston college and International Association for the Evaluation of Eduational Achievement (IEA).

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). TIMSS 2011 Assessment Frameworks.

Thomson, S. Hillman, K., Wernert, N., Schmid, M., Buckley, S., & Munene, A. (2012). Monitoring Australian Year 4 Student achievement internationally: TIMSS and PIRLS 2011. Camberwell, Victoria: Australian Council for Educational Research Ltd.

Hutchison, G., & Schagen, I. (2007). Comparisons between PISA and TIMSS—Are we the man with two watches? In T. Loveless (Ed.), Lessons learned—What international assessments tell us about math achievement. Washington, DC: The Brookings Institution

Martin, M. O., & Mullis, I. V. S., (2006) TIMSS: Purpose and design. In S. J. Howie, & T. Plomp, (Eds.), Contexts of learning mathematics and science—Lessons learned from TIMSS (pp. 17–30). London: Routledge.

TIMSS and PIRLS International Study Center (2011). Informing Educational Policy for Improved Teaching and Learning. Chestnut Hill, MA: Boston College. Retrieved from http://timss.bc.edu/timss2015/participate.html on 30 July 2015

TIMSS and PIRLS International Study Center (1995-2011). Reports on International Achievement in Mathematics and Science, 1995-2011. Chestnut Hill, MA: Boston College. Retrieved from http://timssandpirls.bc.edu/ on 30 July 2015

TIMSS and PIRLS International Study Center (2001-2011). Reports on International Achievement in Reading, 2001-2011. Chestnut Hill, MA: Boston College. Retrieved from http://timssandpirls.bc.edu/ on 30 July 2015

TIMSS Advanced References

Arora, A., Foy, P., Martin, M. O., & Mullis, I. V. S. (2009). TIMSS Advanced 2008 Technical Report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Garden, R. A., Lie, S., Robitaille, D. F., Angell, C., Martin, M. O., Mullis, I. V. S., . . . Arora, A. (2006). TIMSS Advanced 2008 Assessment Frameworks. Chestnut Hill, MA, United States: TIMSS & PIRLS International Study Center.

Mullis, I. V. S. (2014). Introduction: An overview of TIMSS Advanced 2015. In I. V. S. Mullis & M. O. Martin (Eds.), TIMSS Advanced 2015 Assessment Frameworks. Chestnut Hill, MA, United States: TIMSS and PIRLS Study Center, Lynch School of Education Boston College, and International Association for the Evaluation of Educational Achievement (IEA)

Mullis, I. V. S., & Martin, M. O. (2014). TIMSS Advanced 2015 Assessment Frameworks. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Robitaille, D. F., & Foy, P. (2009). TIMSS Advanced 2008 International Report: Findings from IEA's Study of Achievement in Advanced Mathematics and Physics in the Final Year of Secondary School. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

PIRLS references

Joncas, M., & Foy, P. (2011). Sample Design in TIMSS and PIRLS. Retrieved from http://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf on 14 May 2015.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2015). Assessment Design for PIRLS, PIRLS Literacy, and ePIRLS in 2016. In I. V. S. Mullis & M. O. Martin (Eds.), PIRLS Assessment Framework (2nd ed.). Chestnuthill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). PIRLS 2011 Assessment Framework. Chestnut Hill, MA.

Tse, S. K., Lam, W. I., Loh, K. Y., Cheung, W. M., Hui, S. Y. & Ng, H. W. (2012). Progress in International Reading Literacy Study (PIRLS) 2011 International Report: Hong Kong Section. Hong Kong: Faculty of Education, University of Hong Kong. Retrieved from http://www.hku.hk/press/news_detail_8975.html on 30 July 2015.

Other references

ACER. (2013). SEAMEO Experiences of Pirimary Learning Metrics: Desk Review. Retrieved June 17, 2015 from http://seameo.org/index.php?option=com_content&view=article&id=512:southeast-asia-primary-learning-metric-sea-plm&catid=90&Itemid=552

ASER Centre. (2015). Major Research and Assessment Studies in Education. Retrieved May 6, 2015 from http://www.asercentre.org/p/119.html

Bernard, J.-M. (n.d.). Managing the impact of PASEC projects in francophone sub-Saharan Africa.

Betts, J. R. (1999). Returns to Quality of Education. Economics of Education Series 1. The World Bank.

Education Policy and Data Center. (2012). SACMEQ and PASEC. Retrieved from http://www.epdc.org/data-about-epdc-data-epdc-learning-outcomes-data/sacmeq-and-pasec

Chow, K, F., and Kennedy, K,J., 2014, Secondary analysis of large-scale assessment data: An alternative to variable-centred analysis, Educational Research and Evaluation, 20, 469-493.

DeMars, C. (2015). Estimating variance components from sparse data matrices in large-scale educational assessments, Applied Measurement in Education, 28, 1-13.

Ercikan, K., Chen, M. Y., Lyons-Thomas, J., Goodrich, S., & Sandilands, D. (2015). Reading proficiency and comparability of mathematics and science scores for students from English and non-English backgrounds: an international perspective. International Journal of Testing, 15, 153-175.

Froumin, I. & Kuznetzova, M. I. (2012). The Impact of PIRLS in the Russian Federation. In K. Schwippert & J. Lenkeit (Eds.), Studies in International Comparative and Multicultural Education, 13: Progress in Reading Literacy in National and International Context (pp. 183-196). Münster, Germany & New York: Waxmann.

Gove, A., Wetterberg, A. (2011). The Early Grade Reading Assessment: Applications and Interventions to Improve Basic Literacy: RTI Press Publication No. BK-0007-1109. Research Triangle Park, NC: RTI Press. DOI: 10.3768/rtipress.2011.bk.0007.1109

Joncas, M., & Foy, P. (2011). Sample Design in TIMSS and PIRLS.

LAMP Update (2011). Literacy Assessment and Monitoring Programme (LAMP) Update No. 4. Retrieved June 6, 2015 from http://www.uis.unesco.org/literacy/Documents/lamp-update-oct2011-v1-en.pdf

Kebathi , J. (2008). Measuring literacy: The Kenya National Adult Literacy Survey. Adult education and development, Edition 71, DVV international. Retrieved on 10 July 2015 from http://www.iiz-dvv.de/index.php?article_id=802&clang=1

Lewis, S., and Lingard, B., 2015, The multiple effects of international large-scale assessment on education policy and research, Discourse: Studies in the Cultural Politics of Education, 36, 621-637.

Ministry of Education, New Zealand (2011). Statement of Intent 2011/12–2016/17. Wellington: Ministry of Education.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). TIMSS 2011 Assessment Frameworks.

Noble, T., Risebery, A., Suarez, C., Warren, B., & O'Connor, C. (2014). Science assessments and English language learners: validity evidence based on response processes. Applied Measurement in Education, 27, 248-260.

OECD (2008). External evaluation of policy impact of PISA. Paris: OECD.

Plomp, T. (1992). Conceptualizing a Comparative Educational Research Framework. Prospects, 22(3), 278-288.

Robinson, J. P. (2010). The effects of test translation on young English learners' mathematics performance. Educational Researcher, 39 (8), 582-590.

RTI (2014), Early Grade Mathematics Assessment (EGMA) Toolkit. Retrieved May 20, 2015 from www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=157

Rutkowski, L., Gonzalez, E., Joncas, M., and von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting, Educational Researcher, 39, 142-151.

SACMEQ. (2015). Mission.

Save the Children. (2015). Literacy Boost.

SEAMEO (2013). Programmes & Projects. Retrieved May 6, 2015 from http://seameo.org/index.php?option=com_content&view=article&id=512:southeast-asia-primary-learning-metric-sea-plm&catid=90&Itemid=552

UIS (2014), Literacy Assessment and Monitoring Programme. Retrieved June 16, 2015 from http://www.uis.unesco.org/literacy/Pages/lamp-literacy-assessment.aspx

UNESCO (2013). Latin American Laboratory for Assessment of the Quality of Education (LLECE). Retrieved 17 June from http://portal.unesco.org/geography/en/ev.php-URL_ID=7919&URL_DO=DO_TOPIC&URL_SECTION=201.html

UNESCO (2013). The use of student assessment for policy and learning improvement. Retrieved from http://www.unescobkk.org/education/news/article/the-use-of-student-assessment-for-policy-and-learning-improvement/

UNESCO Institute for Statistics. (2014). Learning Metrics Task Force Retrieved June 11, 2015, from http://www.uis.unesco.org/Education/Pages/learning-metrics-task-force.aspx

Unicef (2007). Report of the East Asia Learning Achievement Study. Retrieved June 17, 2015 from http://www.unicef.org/eapro/12205_7702.html

Unicef (2014). Concept Note. Retrieved May 6, 2015 from http://www.unicef.org/supply/files/LRPS_OSR_2014_9112798_Concept_Note.pdf

UNICEF (2014). Statistics and Monitoring, Multiple Indicator Cluster Survey (MICS). Retrieved June 16, 2015 from http://www.unicef.org/statistics/index_24302.html

Unicef/EAPRO, Bangkok (2013). Terms of Reference for consultancy. Retrieved May 16, 2015 from http://www.unicef.org/supply/files/LRPS_OSR_9112798_TOR.pdf

USAID (2013), Early Grade Reading Assessment (EGRA) FAQ. Retrieved May 6, 2015 from file:///C:/Users/susanmarieh/Downloads/EGRA%20FAQs_25Oct11.pdf

USAID (2014), DEP/AME: Philippines Analytic Support Services for Early Grade Reading (PhilEd Data II): Component 2: Early Grade Reading Assessment Results: A cross-language look at MTB-MLE implementation in the Philippines. Retrieved May 17, 2015 from file:///C:/Users/susanmarieh/Downloads/PhilED%20Data%20II%20MT_EGRA_Final_SUBMIT_8-8-14%20(1).pdf

Uwezo (2014). Who We Are. Retrieved June 11, 2015, from http://www.uwezo.net

Young Lives (2015). An International Study of Childhood Poverty. Retrieved June 16, 2015 from http://www.younglives.org.uk/

World Bank Group (2015). READ (Russian Education Aid for Development) Trust Fund. Retrieved June 29, 2015 from http://www.worldbank.org/en/programs/read