United Nations
Educational, Scientific and
Cultural Organization

International Bureau
of Education

In-Progress Reflection No.28

Improving Early-Grade Performance for 2030: Measurement and Estimation Options for Innovative Policy Dialogue

Title	Improving Early-Grade Performance for 2030: Measurement and Estimation Options for Innovative Policy Dialogue
Series	Current and Critical Issues in Curriculum, Learning and Assessment
In-Progress Reflection	March, 2019, No.28 IBE/2019/WP/CD/28
IBE Director	Dr. Mmantsetsa Marope
Coordination and Production Team at the IBE-UNESCO	Carmel Gallagher, Lili Ji, Yi Yang
Author	Helen Abadzi
Keywords	Literacy – numeracy - early-grade reading - early-grade math - growth curves - cognitive science - learning research - SDG 4.1

Improving Early-Grade Performance for 2030: Measurement and Estimation Options for Innovative Policy Dialogue

Abstract

Governments and donors are faced with the challenge of attaining the 2030 Sustainable Development Goals (SDGs). One of the challenges for education is to measure and set benchmarks that indicate whether people have developed sufficient core literacy and numeracy skills to function in the complex literate environments of the 21st century.

Benchmark setting has proved difficult in education because stakeholders find it hard to define which variables really matter and how they can be measured. Measurement of skills particularly in grades 1-3 is highly relevant for policy dialogue, because this is when many students fall behind. Donors use existing national, regional, and international assessments, for which longitudinal data are available. However, assessments require reading fluency and do not focus on the lowest literacy and numeracy levels, so they may overestimate or underestimate learners' skills in grades 1-3.

Governments need specific feedback as soon as possible of likely student performance by 2030 so that they can take measures to improve performance by then. This monograph aims to publicize options for measuring early literacy and numeracy skills, using neuroscientific insights. These may help develop interventions that could accelerate early learning, facilitate monitoring and promote policy interventions to accelerate the achievement of the SDG 4.1 goals in various countries by 2030. The research evidence presented indicates that:

- Performance benchmarks can be set using reading and math fluency research;
- The performance of lower grades worldwide could be monitored through brief tests, measuring concepts that have high predictive validity, derived from cognitive science;
- If these test results were linked to international and national assessments, statistical models could be developing to estimate roughly how populations of various countries are likely to score in 2030;
- Clear feedback, along with recommendations for appropriate interventions, would allow countries and donors to engage in targeted policy dialogue to close the gap;
- Improved performance in grade 1 would improve performance in subsequent grades;
- With sufficient emphasis, funding, and space logistics, illiteracy among low-income students could be eliminated in about 7 years and near-universal and sustainable literacy could be attained by 2030.

Keywords: Literacy – numeracy - early-grade reading - early-grade math - growth curves - cognitive science - learning research - SDG 4.1

Contents

Acronyms

ACER	Australian Council for Educational Research
ASER	Annual Status of Education Report
GAML	Global Alliance to Monitor Learning
EGMA	Early Mathematics Reading Assessment
EGRA	Early Grade Reading Assessment
IBE	UNESCO’s International Bureau of Education
LLECE	Latin American Laboratory for Assessment of the Quality of Education
MICS	Multiple Indicator Cluster Surveys
PASEC	Programme d'Analyse des Systèmes Educatifs de la CONFEMEN
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PPVT	Peabody Picture Vocabulary Test
SACMEQ	Southern Africa Consortium for Monitoring Educational Quality
SDGs	Sustainable Development Goals to 2030
TERCE	Third Regional Comparative and Explanatory Study
TIMSS	Trends in International Mathematics and Science Study
UIS	UNESCO Institute of Statistics
UNESCO	United Nations Educational, Scientific, and Cultural Organization
UNICEF	United Nations International Children's Emergency Fund
UWEZO	Uwezo Annual Learning Assessment

Executive Summary

The challenge

The Millennium Development Goals (MDGs) initiated in 1990 set out a range of international development commitments to meet the needs of the world's poorest, including the aspiration to achieve universal primary enrollment in education by 2015. In 2016, the United Nations launched the bold and transformative 2030 Agenda for Sustainable Development Goals (SDGs) across multiple sectors, with the aim of achieving 17 development goals over the following 15 years. The actions to be taken for the fulfilment of the SDGs involve multiple stakeholders. Education is Target no. 4 which aims to ensure inclusive and equitable quality education and lifelong learning opportunities for all.

The UN 2030 Sustainable Development Goals (SDGs) expect (target 4.1) that:

> (a) *By 2030, xxx% of students worldwide will have attained minimum proficiency in reading and math;*
>
> *(b) By 2030, the proportion of students in country X attaining minimum proficiency will have increased by yyy%, and*
>
> *(c) By 2030, the mean proficiency of students will have increased by zzz% over 2017 levels.*

The 4.1 target could be interpreted as seeking to develop in primary children the minimal skills needed to deal with life in the 21st century. Of particular concern is that the most economically deprived in a country's population can acquire sufficient literacy skills to process the math and reading challenges presented to them.

Purpose of this monograph

The 'custodial agency' responsible for monitoring progress towards SDG4 is UNESCO's Institute of Statistics (UIS). A 'global alliance to monitor learning' (GAML) has been formed to plan and advise UIS. It comprises stakeholders from donor agencies, governments, assessment agencies, universities, nonprofit organizations, and others.

Given the importance of effective monitoring, measurement and early intervention to the achievement of 2030 SDG 4.1., this monograph aims to:

- Disseminate information about innovative measurements that could assess skills easily in the early grades and promote results-based policy interventions to accelerate the achievement of SDG4.1;

- Generate interest, funding and academic collaboration across a range of countries supported by UNESCO, the World Bank and/or other donor and evaluation agencies, to engage researchers in measuring early-grade outcomes in innovative ways, informed by cognitive neuroscience; and
- Use the data from multiple tests to predict likely performance of students in various countries by 2030 and engage countries in a dialogue about closing their performance gaps.

Key Points

1. The limitations of existing assessment instruments

Given funding restrictions and the active interest of many testing agencies, monitoring currently focuses on international and national examinations assessments, for which longitudinal data is already available in various countries (such as PASEC, SACMEQ, TIMSS, PIRLS). With the help of the World Bank, a database has been developed that links learning outcomes across these tests, which makes them comparable, to some extent. Proficiency benchmarks have been developed for these variables. The earliest grade that is assessed by written tests is grade three, although fourth grade is preferred. Such tests do not focus on the lowest literacy and numeracy levels, and items measure slightly different competencies.

International assessments of this type require pupils being tested to be able to read. Multiple issues arise from this strategy, which particularly affect grades 1-3, as many lower-income students in those early grades may be illiterate.

Oral international tests exist, notably EGRA, EGMA, ASER, UWEZO or UNICEF MICS, that ought to measure low performance levels more accurately, but each test measures and defines its levels differently. Some are needlessly redundant (e.g., most EGRA subtests) while others miss critical skills (EGMA). It is difficult to compare results of one test with another. Thus it is very difficult, if not impossible, to determine minimum proficiency in grades 1-3. Tests may overestimate or underestimate learners' skills.

2. Identifying country-level performance projections to 2030 and learning gaps

Test results can be used for statistical modeling and projection of likely performance in 2030. Growth curves can be modeled, such as "growth to standards" models. The percentages of students below a certain benchmark would be estimated, as well as averages of the population and the lower quartile.

Such projections are likely to reveal a gap, and progress may be compared against the default trend to 2030. Thus the scores of a country may be compared against a target and its prior performance, rather than against other countries' performance. Also the percentage of students falling below a benchmark can be compared internationally. (For example, the 2030 EGRA scores may be projected for average and lower-quartile students on the basis of 2006 and 2013 scores. The percentage likely to score below a certain average (e.g., 60 words per minute) in 2030 can be projected). Then Ministries of Education could try to improve instruction to exceed the projections and "move the needle" until 2030.

The methodology is challenging; trends in various countries across time are not always monotonic, and the specifications across tests differ, making equation difficult. However, for some countries, particularly poorer ones, a rough population-based estimate will have value.

3. The importance of processing speed and automaticity

The complex knowledge tested in various subjects through standardized tests rests on a bed of "low level" skills needed in the brain, such as, fluent reading, instant mental math, effortless writing and instant comprehension of long sentences. Attainment of automaticity is a primary goal for all human skills. Specifically for reading, automaticity is the prerequisite skill necessary for all writing systems. After many hours of reading practice, a region of the brain gets activated that recognizes words as if they were faces and processes letters in parallel. Then it is possible for a learner to process volumes of text fast enough for content to be maintained in working memory and be understood. Minimum proficiency can be defined in terms of parallel processing of letters. For math, however, answers must “pop up” effortlessly in the mind.

Practice speeds up performance, so that content can be retained and processed in working memory instantly. Thus, processing speed constitutes evidence of extensive prior instruction, practice and education quality in all topics.

In the early years, speed is a crucial variable. As speed increases, its importance diminishes. For this reason, first graders should at least learn how to decode transparent orthographies, vocabulary needed for class operation in official languages (if needed), and at least single digit operations (with approximate math and operations on a number line). Grades 2-3 should increase speed and accuracy. To meet any benchmarks in grade 3, students should become proficient in nearly all competencies in the curricular frameworks of grades 1-2. Otherwise, it is impossible to become proficient in subsequent ones.

Performance prior to automaticity is not comparable across languages and scripts.[1] At the early levels, performance and progress would be compared to curricular standards of countries (e.g. no. of Chinese characters mastered, knowledge of matras and conjoint consonants of Indian scripts, decoding of irregularly spelled patterns in Khmer, etc.). By the end of grade 3, essential

[1] The languages and writing systems of the world seem too dissimilar for comparison, but they constitute inputs into the same brain regions. The analogy is electric chargers of varying voltages that ultimately power the same computer motherboard (Chapter 2). Words per minute also appear non-comparable, but there are multiple methods for equating them (Abadzi, 2013; Chapter 3).

reading instruction would be complete in almost all writing systems. Then words-per-minute performance can be set for automaticity in various languages and scripts.

Minimum proficiency can be defined in terms of the speed necessary to perform essential tasks. For example, participating countries by 2030 could aim to have the vast majority of their students who finish grade 3, (a) read at least 60 correct words per minute and (b) make at least 20 correct single-digit calculations per minute. (Cross-language metrics exist.) At these speeds, performance is probably automatic. Even if citizens are slow, they may have sufficient working memory available to make sense of content. Their performance would be very different from citizens who know letters but must make conscious effort in decoding or counting.

If students become fluent in school and are taught well, they will score high in the early-grade skills. These variables also predict future performance. For example, a reading speed of 150 words per minute, a speed of 50 single digit operations correct per minute, a high score on language tests bear witness to considerable schooling and some confidence that a student will deal with relevant information reasonably in life. In case students drop out in primary school with these competencies, they may apply them fluently to their daily circumstances and thus improve. This is what some research suggests (e.g. Hartley & Swanson, 1986).

4. Use of cognitive science to define and monitor the SDG target for grades 1-3

As indicated above, stakeholders involved in SDG deliberations have found it hard to define minimum proficiency at various levels. Cognitive science and memory can provide a framework to gauge and compare minimum proficiency. The rationale is as follows:

- Despite cultures and individual variation, there are robust neurocognitive processes that are common among people; in fact they account for the curricular similarities worldwide and the ability to communicate with other humans on the same topics. Using these concepts the assessment tasks can be redefined to assess mental commonalities and could improve accuracy and predictive validity of measure in the low grades.

- It is not sufficient for students to "know" something. They must be able to retrieve it in milliseconds and connect it to other instantly retrieved items in order to answer problems and make decisions. The time available to do this is determined by working memory, which holds perhaps 7 items for about 12 seconds. Thus students must process information as fast as possible. Processing speed increases when students spend time practicing tasks. Practice automatizes execution, so students do not have to think of every step. If they perform reading or math tasks instantly and effortlessly, they can focus on inferences and complex problems.

- Transparent orthographies can be taught quickly, some just in the first half of grade 1. Practice then creates automaticity, and the first graders of some countries can become basically literate in one semester. Also remediation groups can be formed and do the same for an entire school. Improved performance in grade 1 would improve performance in subsequent grades.

- With sufficient emphasis, funding, and space logistics, Illiteracy among low-income students may be eliminated in about 7 years. Near-universal and sustainable literacy is really much closer than commonly thought and could be attained by 2030.

5. The importance of monitoring essential, automatized skills

UIS decisions about monitoring have been complicated by what appear to be 'common sense' ideas about learning. Some of the concepts being discussed here are obscure. The relevant variables are beyond our awareness, and it takes specialized training to understand them. Since "low level" skills are unconscious, national curricular frameworks often neglect them. For example, an IBE analysis of math curricula demonstrated the neglect of "proficiency" assessment (Chapter 3). This memory bias against unconscious processes also creates an issue of "face validity." People find it hard to believe that processing speed matters so much, so they focus on comprehension and complex reasoning. But these preliminary skills must be attained in order for complex cognition to take place. Without a minimum processing speed, inferences and critical thinking are impossible.

As a result, some stakeholders define reading only as comprehension, skipping the crucial perceptual learning and decoding stages. However, comprehension is not a single variable (Chapter 2). Speed matters, which partly depends on language command. Therefore, early-grade comprehension can certainly be measured, but language knowledge and speed of comprehension may be more important.

It is important to monitor performance in these "low-level" skills, not only for their own sake, but because attainment suggests good performance in more advanced subjects. In conjunction with performance speed, certain early competencies have high predictive validity for the performance of higher grades. It is thus possible to set early benchmarks for language, spatial reasoning, and numeracy.

6. Options for monitoring early literacy

UIS and stakeholders must make the best use of existing tests, and none of these consider processing speed explicitly. There are some options for doing so.

- Students taking international tests could also take one-minute reading and calculation tests that monitor speed and secondarily comprehension.
- It may also be possible to use proxies for speed. For example, in citizen tests, enumerators may state their opinion on a reader's automaticity, and previous works suggests that it may be reliable.

- Latency is another option, e.g., the time that elapses before someone responds. (Test-taking through tablets could measure latencies.) In surveys, participants may be asked if they read messages on television or store signs; these would be indicators of automaticity.
- With some research, a model may be developed to translate roughly the various oral tests into word per minute equivalents. Then it may be possible to measure, compare, and monitor across time the performance at the lowest levels. Such work will also be relevant to assessments of adult literacy.
- Reading speed and accuracy have been extensively measured in more than 70 countries through the Early Grade Reading Assessment (EGRA). This is an extensive test, of which only 2-3 subtests have high predictive validity (notably 1-minute reading speed and comprehension questions). EGRA is individually administered, but the 1-minute connected reading test may only take 5 minutes or so. (A few group test alternatives such as Wordchains are possible in some languages.) The texts do not have to be internationally equated. They would pertain to texts similar to those found in grade 2 textbooks. Therefore these are regarded as curricular assessments in the US. In some countries and languages, scores can be equated with PASEC (grade 2) and with PIRLS, grade 4.
- For grades 4 and above, PIRLS and other international tests are to be used (Progress in International Reading Literacy Study). To perform satisfactorily, 4th graders must read passages of 800-1000 words and answer a set of short-answer and multiple-choice questions in 20 minutes. This necessitates a rate of about 100 words per minute through 2-3 pages. "Citizen" assessments such as ASER, Uwezo, or the UNICEF MICS are to be used, but time-bound performance must be somehow established for fluency. Estimated words per minute may give a clearer view of what citizens are able to perform in terms of daily tasks.

Most countries in the world use multiple languages, sometimes with religious implications. Citizens should be fluent in the grammar and writing system used officially, not only their own. Measurement planning and budgeting in each country must take that into account.

7. Options for monitoring early numeracy

Numeracy is innate in the brain, and students learn to extend their inborn system to higher levels. Working memory is a crucial issue, as with reading. Therefore there is a need for fluent and automatic math performance that instantly pops into students minds (e.g. 57+8). In addition there are variables related to the numbers line, estimations, fractions comparisons and digits correctly operated per minute that highly predict math performance after the early grades. Particularly important is cardinality that is the quantity represented by number symbols. Cardinality speeds up numeracy acquisition, and research suggests that this should be done by age 4. Attainment therefore could function as a benchmark for preschool (Chapter 3).

The Research Triangle Institute developed EGMA, the Early Grade Math Assessment that has been given in many countries. This is very useful, and data can continue to be collected. However, it must be augmented with per-minute calculation tasks and potentially with estimation tasks. One alternative that is administered in groups is the Number Sets Test (Geary et al., 2007, in the US). Research on its concepts is extensive and state-of-the-art. This test could be adapted for international use and enhanced for higher performance levels also.

The most likely combination of math tasks with the highest predictive validity of long-term outcomes would involve:

(a) Fluency in solving basic addition and subtraction problems; e.g., how many items can be solved correctly in 60 sec.

(b) Understanding of the relative quantity of numerals, e.g., which is larger 42 or 29; for first graders values should be smaller;

(c) Fluency of accessing quantities represented by numerals. This could be the Number Sets Test or a timed test on identifying the larger of two small numerals e.g., 1 vs. 3, or 2 vs. 9.

All of these will be correlated with one another, and composite scores should be predictive of long-term outcomes in math, controlling other factors (e.g., Geary et al., 2013).

There is considerable educational research focused on predicting student performance on the basis of information-processing variables. The performance of lower grades worldwide could be monitored through: (a) brief and cheap tests measuring concepts that have high predictive validity, derived from cognitive science; (b) Linking test results to international and national assessments; and (c) developing statistical models to estimate roughly how populations are likely to score in 2030, then engaging in policy dialogue to close the gap. If funding were available, a measurement plan and strategy would include steps outlined in a subsequent section.

8. Minimum proficiency for the early grades

The core memory commonalities facilitate the estimation of parameters that have been hard to define. Target 4.1 could ensure that every child, regardless of circumstance, completes primary education able to read, write and count well enough to meet minimum learning standards. A methodology for determining these is in Chapter 4. It uses the percentages of students deemed competent under a normal curve for various countries and may determine how to increase performance for countries that score 1 or 2 standard deviations below the mean in a specific measure.

- By 2030, 83% of students worldwide will have attained minimum proficiency in reading and math

- By 2030, the proportion of students in country X attaining minimum proficiency will have increased by xxx% or by 0.20 standard deviations (projecting from earlier repeated tests).

- By 2030, the mean proficiency of students will have increased by zzz or by 10 words per minute over 2017 levels (projecting from earlier repeated tests).

Neurocognitive research can help set country-level benchmarks. By 2030, participating countries could aim to have the vast majority of their students who finish grade 3, transition from 'learning to read' into 'reading to learn'. This means: (a) read at least 60 correct words per minute and (b) make at least 20 correct single-digit calculations per minute. Lower-income countries could aspire to educate students who score 1 standard deviation below the mean, i.e. attain the above benchmarks for about 83% of the students (calculated as area under a normal curve). Better-off countries could aspire to attain these benchmarks for students scoring 2 standard deviations below the mean (i.e. about 97% of a normal curve). These goals would exclude special education students.

A parallel or alternative method is to "teach to the test" standards of PIRLS and TIMSS. Countries may aspire to attain a mean of 500 in these tests by 2030. In each case certain competencies are needed. Using early-grade tests, countries can assess the attainment of the prerequisites that will make the score of 500 attainable. For example, to read a passage of 1000 words in grade 4 at about 100 words per minute, the vast majority of students should become fluent readers at 60 words per minute perhaps by the middle of grade 2. Practice aimed at fluent performance of more complex skills in the early grades should be specifically included in curricula.

9. The need for applied research funding and knowledge exchange

Many of the concepts discussed in this document have been developed in academic settings across various countries, such as the US, Canada, Spain and China. Additional thinking is needed to streamline and compare one-minute reading tests, as well as linking them to international tests for higher grades. For example,

- The psychometric properties of speed tests need to be revisited, given that speed tests have been neglected in recent decades,
- Words per minute must be compared across languages, and
- The utility of growth curves needs to be tested.

Chapter 1. Background: Curricular variability and comparison difficulties across countries

The UN initiative on the 2030 Sustainable Development Goals (SDGs) expects for all children (target 4.1; UNESCO, 2015) that:

- *By 2030, xxx% of students worldwide will have attained minimum proficiency in reading and math;*
- *By 2020, the proportion of students in country X attaining minimum proficiency will have increased by yyy%;*
- *By 2020, the mean proficiency of students will have increased by zzz% over 2017 levels.*

The governments and other stakeholder organizations have been expected to estimate and report on intermediate figures periodically. How can stakeholders know the extent to which SDG targets have been achieved? But how to substitute the above xxx's with real numbers? How can stakeholders know that SDG targets have been achieved? What is the most parsimonious way to measure progress across countries and across time? These questions have been hard to answer.

The development of methods to monitor the fulfilment of the SDG goal 4 is a complex process in all sectors, including education. Governments, who act in a voluntary role, are expected to meet certain performance targets by 2030 and demonstrate progress until then. The custodial agency for this target is UNESCO through its Institute of Statistics (UIS). The agency must report progress at regular intervals to the United Nation General Assembly and also report on the reliability of the indicators used. To decide which indicators to use, task forces have been created. Dozens of organizations have some level of involvement in these task forces, and participants have varying levels of expertise in learning or measurement.

In 2017-18, efforts were made to compare the curricular frameworks of reading and math and to place various competencies at specific progression levels. The process proved laborious and showed that curricular frameworks have limited comparability, even in the lower grades. A more promising means of comparing learning outcomes has been through standardized international tests using various schedules to different combinations of countries, grades, and languages. The World Bank supported an ambitious and complex statistical methodology to harmonize learning outcomes and thus create equivalencies among the many tests (Altinok et al., 2018). Efforts in the future may include various national tests. Thus, monitoring efforts in various countries are expected to rely on assessments that already have data available.

A study by Treviño & Órdenes in 2017 explored the commonalities and differences between regional and international assessments, to understand the challenges and options for reporting on progress towards indicator 4.1.1. The analysis suggests that the different approaches to measuring indicator 4.1.1 have advantages and shortcomings in relation to technical issues and feasibility. Given the diversity of assessment programs it is necessary to establish political and technical agreement on the minimum level of competency in reading and mathematics. It is also necessary to promote procedural consistency to ensure minimum data quality. The paper proposes four strategies for reporting indicator 4.1.1, including a new and unique SDG4 test.

By the end of 2018, following many consultations, the task force agreed on global indicators and made significant progress on the organizational and conceptual work for assessment and. UIS also settled on a portfolio approach to monitor outcomes across countries. These include:

- curricular, non-statistical approaches: progression levels and expert judgments to determine details:
- statistical approaches consisting of:
  (a) item-based linking, psychometrically informed recalibration based on common items taken by different individuals,
  (b) test-based linking: recalibration based on giving parallel tests to a representative sample of respondents,
  (c) statistical alignment: recalibration of existing data using countries who participated in more than one cross-national assessments (a "Rosetta stone" approach).

These efforts have been partially effective in aligning countries. Various governments monitor multiple variables that cannot be easily compared and attempts to create equivalencies show stark differences in performance between higher and lower income countries that cannot be easily breached. In some cases, performance at the 50th percentile of OECD countries corresponds to performance at the top 5% of students in poorer countries.

To facilitate country reporting, UIS has developed a content alignment tool and a procedural alignment tool to help countries to align their national assessment results with a global reporting scale. Content alignment means that a country with national assessment in grades 2/3 should decide: is my math or reading assessment sufficiently aligned with what is meant by math or reading at a particular indicator level? For countries that agree, UIS has developed an online tool that will receive the data and generate patterns regarding the extent to which countries align to a universal common framework that makes it possible to report against a global indicator. Very generally, in math the areas that are being aligned include: number knowledge, measurement, statistics, geometry, and algebra. (A determination of "math proficiency" has been excluded due to insufficient agreement on content coverage). For reading, the content areas include: reading competency, linguistic competency, and metalinguistic competency.

Theoretically, there are four assessments that measure proficiency in Grades 2 and 3 (EGMA, EGRA, LLECE and PASEC). Harmonized learning outcomes certainly have much explanatory and monitoring potential. However, it is unclear how to test grades 1-3 systematically, particularly in low-income countries, and in the Arab world, where many are still illiterate. Some of the reasons are:

- **Paper-and-pencil tests are applied too late to help students by diagnosing problems and facilitating targeted interventions**. The multiple-choice assessments typically start in grade 4, when written tests can be administered to groups. However, student failure in basic competencies starts before grade 1 and is exacerbated if it is not address early in formal schooling. Many tests, including the many Early Grade Reading Assessments (EGRA) and Early Grade Math Assessments (EGMA) show widespread illiteracy and ignorance of numerical procedures.
- **Unrealistic expectations of rapid progression**. Many experts who developed coding schemes and progression scales are from middle-income countries, whose students quickly master the basics. The curricular milestones of grades 1-3 in many countries reflect the middle-class and middle-income orientation of the writers, where early performance quickly improves. For example, content areas of alignment in reading start with reading competency and move on to complex linguistic competencies, neglecting the processes of learning to read. Since the goal is basic education for the entire population, proposed levels and progression scales may not realistically reflect population performance.
- **Exclusive focus on monitoring with limited attention to instructional interventions**. Assessment programs have been carried out for 20 years, and countries have received much training and financing to test their students. Yet, few seem to have benefited from feedback. There are signs of test fatigue, and some officials raise questions about the value of more testing. GAML could be more attuned to instruction, but its purpose is monitoring, not instruction.
- **Financial and institutional uncertainties**. UIS and other stakeholders estimate that roughly US$250 million will be needed for the planned testing program, along with much international and national capacity building for measurement. Donors in 2018 were reluctant to finance large measurement initiatives. Funding and institutional uncertainties make it difficult to hire top researchers with many publications in the various areas.

Yet, reading and math achievement in the early grades must be measured, and not just for the sake of these subjects. Performance forms the basis of information processing of all subjects, from history to science. It is important therefore to have benchmarks that signal the desirable performance level for current and future attainments.

Needed: A few highly predictive variables for worldwide achievement monitoring. Though languages and scripts differ superficially, they are executed by the genetic mechanisms that are common to all humans. Therefore, the solution may lie not in comparing the math curricula or the writing systems per se, but the amount of information students' brains can process when they are at various points of development.

***Using the human cognitive commonalities to monitor target* 4.1.1**

The following section outlines the memory systems used by students' brains, which will help to identify the tasks that must be tested.

A. Memory functions lead to global benchmarks

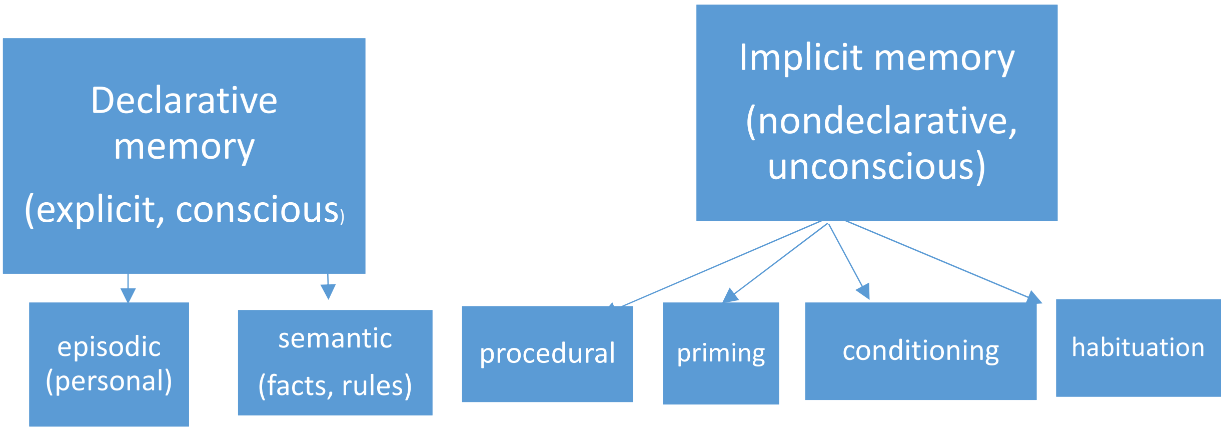
Memory exists to help organisms survive and has an architecture that is common across humans and also other animals. Learning involves the development and strengthening of neural connections in response to various stimuli. From an educational aspect, learning also consists of linking small units into larger units through practice. To learn academic and vocational skills, humans must piggyback on this ancient system. Memory functions help explain what is retained and what must be taught and tested. The underlying cognitive and perceptual commonalities make it possible to compare performance across countries, albeit roughly. Shared ways of thinking are the reason for curricular similarities worldwide, despite cultural differences.

Humans obviously have individual differences in mental processes. Some of us may excel in features such as spatial reasoning and perception that assist some to do better drawings for example. But schooling should enable everyone to reach a minimal competency in skills that predict performance of future tasks. What schools should produce and the 4.1 target should monitor is not only knowledge of single items and procedures but also performance that is fluent and overlearned (therefore relatively resistant to forgetting) in these core competencies.

Types of memory and their implications for target 4.1

Very roughly, our long-term memory is divided into two systems: (a) explicit or declarative conscious recollections of events and facts and (b) implicit memory, which allows us to store instructions on how to do things. Declarative and implicit knowledge are closely linked and complementary. Declarative memory involves conscious "knowing that…" whereas implicit memory involves knowing "how to..".

Figure 1.1: Memory types (Squire, 2004)



Within declarative memory, personal recollections are called episodic memory, while conscious knowledge of facts is called semantic memory. Within implicit memory there is memory for procedures (procedural memory), priming, conditioned responses, and habituation to the environment. Social learning and adaptive imitation also form parts of procedural memory.

Because implicit memory is unconscious, people cannot easily talk about it. It can be thought of as a dark network, whereas declarative memory is exposed to the sunlight of conscious thought. The duality is poorly understood, and many educational misunderstandings result from it.

Skills must be "compiled" for automaticity and speed. Schools most easily teach facts and rules, using books, technology, classroom activities and discussions. Facts constitute conscious, declarative memory. They may be pulled out of declarative memory and become compiled into chains that can be quickly retrieved. Practice helps transform these chains of knowledge items into sequences of automatized actions. Then they are stored in the implicit memory system of the brain, and many can be retrieved at will. Application of facts and rules learned in school creates knowledge, skills and "experience".

Working memory: Our brains link information and reach decisions in a space called "working memory." It contains the items our minds process at a particular moment. But that space is very limited; it holds about 4 to 7 information items for 12-25 seconds (Baddeley et al. 2015).[2] This leaves little or no time to think consciously about everything that we do. We must make a few key decisions and let many other tasks run unconsciously. The way to overcome the working memory bottleneck is to pass long chains of actions through it. The chains may then count as one item. This privileges acts or thoughts that are performed in milliseconds (Cooper & Sweller, 1987). If some components take longer, people may forget what they were doing; they may get confused, lose patience, and eventually stop.

To fit a message into working memory, speed is necessary. Speed is attained by practicing extensively small units, which are linked into larger changes and rearranged in more efficient sequences. Practice also takes the action out of conscious thinking. After many trials, content transfers to implicit memory and bypasses working memory. This leaves mental space for comprehension and complex thinking, while a student is also doing something else, such as taking notes or calculating. If speed does not increase, people must think consciously about every move, such as finding each keyboard key to type. Therefore ability to execute a chain of actions fluently, and with little conscious thought, constitutes evidence of practice or training. Processing speed of various tasks, measured in seconds and milliseconds, should be an essential benchmark of

[2] See a popularized blog: https://www.ukfiet.org/2017/sdgs-social-equity-and-the-contents-of-your-working-memory/

specific skills. Milliseconds seem insignificant, but they accumulate, and delays eventually overwhelm working memory.[3]

Young learners most easily learn skills that rely heavily on implicit memory (Rovee-Collier et al., 2000.) For example, students may easily learn acrobatics, music, horseback riding, roller skating, animal husbandry, field cultivation. The maturation of the brain over time facilitates access to explicit memory, reasoning, and complex cognition, as well as comprehension and the acquisition of more "cognitive" skills. But advancing age gradually decreases the ease of entering material into implicit memory. Thus, lower-income students may be quite competent in tasks that rely heavily on procedural memory. But due to limited educational opportunities and possibly also developmental, nutritional, and cultural reasons, they may face delays in the maturation of explicit long-term memory and attention span. They may read fluently but understand poorly.

Grade- and age-based benchmarks thus give different information for poorer populations. On one hand, certain skills are taught in certain grades, and students must attend to them in order to learn specific content. On the other, age matters for comprehension and reasoning. Some educators are impressed by older students' ability to learn in a few months the curricula of the lower grades. However, overage children have little time left for schooling, particularly if they are girls. They may have to drop out for work or marriage. The skills attained by 9-year old first graders therefore, may reflect the skills they will attain upon leaving school. This is why data on age and grade are needed to clarify attainment.

Since the memory requirements for effortless performance are unknown, they are often misunderstood and neglected. Particularly striking is the lack of curricular attention to low-level skills that must be automatized (for example, proficiency objectives in math). In the preliminary IBE study of National Assessment frameworks,[4] only 10-30% of countries had objectives related to 'proficiency'. Curricula showed a justifiable focus on necessary procedures, but without establishing a minimum processing speed, comprehension and complex thinking are impossible. Without stringing foundation units together, students may expend effort but process little text, and they may forget content in the long run.

Many countries may not allocate in curricula sufficient time or instruction to developing processing fluency in various tasks, particularly among poorer or lower-scoring students.[5] Some

[3] Some people are faster than others in executing various skills, so processing speed is an element of intelligence; scores for both the Working Memory and processing Speed subtests make up the WISC-IV's Cognitive Proficiency Index. However, training improves everyone's performance.

[4] https://teams.unesco.org/projects/gal; http://uis.unesco.org/en/topic/learning-outcomes

[5] For example, in the 2006 Kenyan syllabus, fluency goals were merely inferred: by Standard 2 children are expected to "read simple sentences/passages related to greetings and polite language" (Objective 1.2.d), colors (2.2.f), numbers (4.2.e), time (5.2.e), position and direction (6.2.e), home and home activities (7.2.e), shopping (8.2.c), their bodies (9.2.e), health and hygiene (10.2.c), travel (11.2.f), clothes (12.2.c), food (13.2.d), wild animals (14.2.c), weather (15.2.c), the farm (16.2.c.) and home equipment (17.2.c. and d.).

evidence may be derived from the percentages of students deemed proficient in fundamental competencies of various domains.

Students' working memory is often assessed in educational research, but rarely in international education studies.[6] Since answers to reading math questions depend on working memory capacity and speed of processing, working memory assessments could, and should be included in test batteries (Geary et al., 2009). Such tests are usually brief and can provide insights into students' ability to retrieve and briefly retain information items. One recently popular test is the N-back.[7]

B. What is Proficiency? Fluent and relatively effortless performance

As discussed earlier, practice creates automaticity and frees up working memory to be engaged with the more conceptual aspects of a task. Inordinately long and effortful tasks create brain fatigue and tend to get abandoned (e.g. Mizuno et al. 2011). Students cannot afford to stop and think hard for every procedure they execute. Laborious execution suggests that the competency will soon be forgotten (Baddeley et al. 2015 on overlearning and consolidation). Two students may obtain the same score in an exam, but a month later, the student who has practiced extensively and is fluent may remember more than the other.

Memory functions help define "proficiency". The term proficiency implies not just conscious recall of some information, but automaticity and fluency in execution; that is a delay of only milliseconds between retrieving one unit in a procedure and the next. It also implies that improvement has reached a plateau in a learning curve.

[6] The Working Memory Test Battery for Children (WMTB-C; Pickering & Gathercole, 2001) consists of 9 subtests that assess the central executive, phonological loop, and visuospatial sketchpad. All subtests have 6 items at span levels ranging from one to 6 to 1 to 9. Passing four items at a level moves the child to the next level. At each level, the number of items (e.g., words) to be remembered is increased by one. Failing three terminates the subtest.

[7] The **central executive** of the working memory is assessed using three dual-task subtests. Listening Recall requires the child to determine if a sentence is true or false and then recall the last word in a series of sentences. Counting Recall requires the child to count a set of four, five, six, or seven dots on a card and then recall the number of counted dots at the end of a series of cards. Backward Digit Recall is a standard backward digit span.

Phonological loop. Digit Recall, Word List Recall, and Nonword List Recall are standard span tasks with variant stimuli; the child's task is to repeat words spoken by the experimenter in the same order as presented by the experimenter. In the Word List Matching task, a series of words, beginning with two words and adding one word at each successive level, is presented to the child. The same words, possibly in a different order, are then presented again, and the child's task is to determine if the second list is in the same or different order than the first list.

Visuospatial sketch pad. Block Recall is another span task, but the stimuli consist of a board with nine raised blocks in what appears to the child as a "random" arrangement. The blocks have numbers on one side that can only be seen from the experimenter's perspective. The experimenter taps a block (or series of blocks), and the child's task is to duplicate the tapping in the same order as presented by the experimenter. In the Mazes Memory task, the child is presented a maze with more than one solution and a picture of an identical maze with a path drawn for one solution. The picture is removed and the child's task is to duplicate the path in the response booklet. At each level, the mazes get larger by one wall (Geary et al., 2009).

Automaticity requires practice and guidance, but proficiency should not be assumed even in higher-income environments. Many schools focus on enjoyment and creativity, so students may not become fluent in writing or in math calculations for years (if ever). It is thus uncertain that the proportion of students attaining minimum proficiency in reading or math will increase by 2030, unless specific interventions are actually implemented. In fact, students may stagnate over time in some fundamental competencies, such as mental math.[8]

The simplest way to determine *minimum proficiency* would be to establish correct items per minute for various grades and subjects (for math, see Fuchs et al., 1998). This is obviously easier for the lower grades and more challenging for content that also requires execution of multiple chains, as well as inferences. For more complex cognition, proficiency benchmarks have been developed in existing international and regional studies (e.g. classifications above 625 in TIMSS). Attainment of high benchmarks suggests that thousands of hours have been spent in practice and instruction.

Stakeholders use some words informally and imprecisely, such as "children are not learning". Memory research provides greater nuances in understanding what this may mean. Children "not learning" may mean falling below minimum specifications in cross national assessment. The children may lack the cognitive networks to attach new information in the higher grades, but they may process information about life-related issues. If they do not know algebra they may not learn quadratic equations, but may be able to do reasonable arithmetic. It is useful to use terminology precisely.

Figure 1.1: IBE document – National assessment frameworks

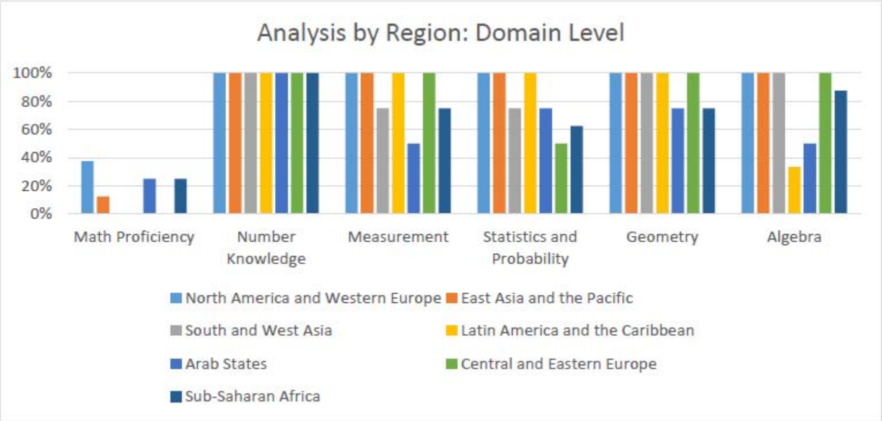


Figure 8: Analysis by region: Domain level

[8] Mental math examples are: John, 12-years-old, is three times as old as his brother. How old will John be when he is twice as old as his brother? Or: Two families go bowling. While they are bowling, they order a pizza for £12, six sodas for £1.25 each, and two large buckets of popcorn for £10.86. If they are going to split the bill between the families, how much does each family owe? Or: 4, 9, 16, 25, 36, ?, 64. What number is missing from the sequence? (https://www.tests.com/practice/WISC-Practice-Test)

Because proficiency is poorly understood, instruction and testing tends to neglect it. A UNESCO-IBE study on national assessment frameworks for math showed little measurement attention towards "proficiency" (however specialists of each country defined it). Curricula in all regions focused on number knowledge, but only 0-40% focused on proficient execution of relevant skills. What is not tested is often not taught, so this may be one reason why opportunities for practice to speed up execution and automaticity may be decreasing in schools.

Emphasis on automaticity for basic process resolves some debates regarding grades 1-3. Given the shared human DNA that creates memory functions in the brain, it is possible to use speed and fluency scores to set minimum proficiency benchmarks worldwide. Tests rather than items can be set, and these can be national or regional (determined by language and script). Equating or using the same reading items is unnecessary, given the script and language variations.[9] Processing speed is the initial main variable. (Numeracy is innate and is typically similar across countries.) Countries can be advised to focus on these early proficiencies in order to achieve more advanced ones. It is important to layer skills, starting with basic skills, in order to attain complex cognition. Decisions to create indicators without automaticity considerations are akin to 21st century doctors deciding that they will not use antibiotics.

[9] Equating refers to procedures for making assessments comparable, either among different forms or over time.

Chapter 2. Foundational reading: The role of perceptual learning and increasing speed

Many studies suggest that the reading network in the brain is controlled by the organization of the network pertaining to speech. Across a wide spectrum of languages, speech–print convergence is a common signature of reading proficiency. This happens whether the writing system is alphabetic or logographic, whether it is opaque or transparent, and regardless of the phonological and morphological structure it represents. (Rueckl et al, 2012).

Therefore people worldwide, including the blind, read using the same parts of the brain and rely on processes that are universal (e.g., Perfetti et al. 2013). Reading automaticity is a feature common to all brains and measurable through event-related potentials; but the roads leading to it can be simpler or more tortuous. Some reading systems are more complex than others, so attaining fluency requires varying timeframes. Deep orthographies like English take years, while transparent orthographies with few letters, like Dhivehi, can be automatized in less than a year. Scripts with greater perimetric complexity need more practice for automaticity, even if they are transparent (e.g. Hindi, Arabic; see work by Chang, Chen, & Perfetti, 2017; Shimojo and Chengizi, 2005).

To find common parameters in all languages and script, the face recognition mode of reading in a certain script is an essential goal to attain worldwide. In all but about 9 languages (including English and French), decoding should be mastered by the end of grade 1. For the syllabic scripts of India and Ethiopia, two grades should be necessary in order to learn the conjoined consonants (and Ethiopian fidel have some unpredictable combinations). For mastery of essential Chinese characters about 4 years are needed. Thus grades 2-3 should increase speed and accuracy at least of the 3 Rs (and to learn most combinations of English and French). To meet any benchmarks in grade 3, students should become proficient in nearly all competencies prescribed in the curricular frameworks of grades 1-2 for their script. Otherwise, it is impossible to transition to “read in order to learn”.

The ability to perceive instantly the symbols of a country’s script(s) is a crucial predictor of future performance. Initially letter-by-letter recognition and mapping to sounds is necessary. Practice links small units into larger ones. The input passes through two tunnels, a visual and a working memory tunnel. To be understood, a message must stay there long enough for information to rise from long-term memory and interpret it (if students know the language). Letter-letter-reading is too slow for working memory to contain a message. Instant reading and comprehension become possible only when an area is activated that processes letters in parallel and sees words as if they were faces. “Face” recognition comes first, comprehension necessarily comes later (See Abadzi 2017 for details and citations.) Minimum proficiency can be the attainment of parallel processing.

As students practice, the conscious recall of letters becomes automatic, bypasses working memory, and leaves most of its space available for interpretation. This stage may be happening at 45-60 words per minute, depending on word-counting methods. At around 45-60 words per minute the visual word form area is consistently activated, which recognizes words as if they were faces (Dehaene & Cohen, 2011). Only then can students read with sufficient attention to content. Thus reading speed matters in most real-world contexts, and it is a robust and easy aspect of reading to measure. It suggests that certain neuronal groups in the brain have been wired and fired together thousands of times. Theories of reading should account for speed (Pelli et al., 2012).

The reading function reuses neurons that evolved for tracking small objects in space and is not innate. Thus reading completely depends on a *perceptual learning function* (Yu et al., 2009). Some studies have compared the perceptual learning variables applicable in various languages and scripts. For example, a measure of graphic complexity in 131 languages (Chang, Chen, and Perfetti, 2017) can be universally applied across writing systems, providing a research tool for studies of reading and writing.[10] Under all circumstances the brain expects rapid and accurate input in order to send input into the comprehension areas. This is what can be measured.

The following figure gives an example of visual complexity parameters. More complex shapes will take longer to automatize than scripts of simpler shapes, but automaticity should level the ground among scripts.

Figure 2.1: Perimetric complexity parameters (Chang et al. 2017)

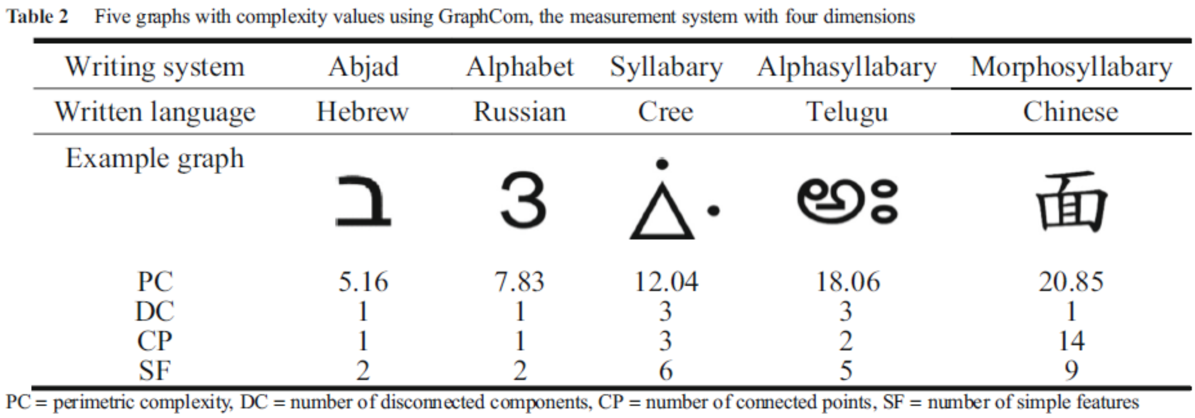


Table 2 Five graphs with complexity values using GraphCom, the measurement system with four dimensions

Writing system	Abjad	Alphabet	Syllabary	Alphasyllabary	Morphosyllabary
Written language	Hebrew	Russian	Cree	Telugu	Chinese
Example graph	ב	З	ᐄᐧ	ఱః	面
PC	5.16	7.83	12.04	18.06	20.85
DC	1	1	3	3	1
CP	1	1	3	2	14
SF	2	2	6	5	9

PC = perimetric complexity, DC = number of disconnected components, CP = number of connected points, SF = number of simple features

As indicated in the section on memory basics, further practice increases speed and gradually takes the deciphering process out of the conscious memory and into implicit memory. This enables working memory to attend to the meaning, while the visual and decoding aspects run in the background. Reading speed, therefore, is a highly predictive variable of future performance, even if a student does not know a language well at that time. The research implies two broad stages of reading acquisition: one before automaticity and one after. Obviously the latter stage has multiple gradations, as the reading regions of the brain become better connected. Speed

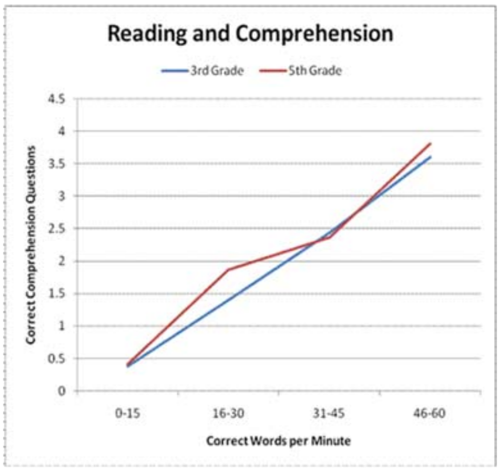
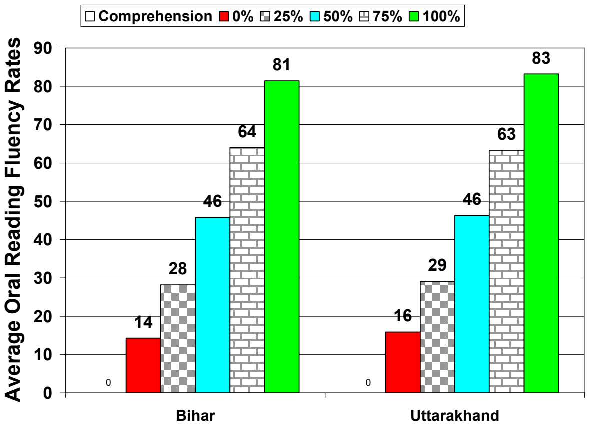
[10] GraphCom's 4 dimensions are: number of connected points, number of disconnected components, number of simple features, and perimetric complexity. A composite score represents the overall complexity of a grapheme in a writing system. Also see Chapter 5.

and prosody improve, actual vocabulary and grammar knowledge as well as effortless access to it improve. But speed limits the rate at which information is processed by the reader. When decoding speed cannot rise, the reader experiences a disability. The range of print sizes that maximize reading speed is highly correlated with the character sizes used in printed materials (Pelli, Chung, & Legge, 2012).[11]

Comprehension. Expert readers instantly access language when they read, so language stands out as the main feature of reading. But reading is not language.

Comprehension constitutes retrieval of meaning from long-term memory. It is not a single variable; it is a complex concept with multiple ramifications. It has automatic as well as conscious aspects. It depends on speed and accuracy of concept linkages, and for that vocabulary is insufficient. Instant access to grammar rules is needed. Effortless comprehension is crucial; fuzzy and slow processing makes people run out of working memory. And if they must think hard, they certainly do not enjoy reading. Fuzzy and slower language recall impacts many countries that have 'diglossia' (a term applied to languages with distinct 'high' and 'low' colloquial varieties, such as the Arab world and German-speaking Switzerland).

Figure 2.2: The relation between fluency and comprehension

Ghana	*Samples in two Indian states, grades 1-5*
	
To answer 80% of the questions correctly, students had to read more than 60 words per minute (RTI 2009)	2nd grade test in Devanagari script read by grades 1-5 shows the relationship between speed and comprehension (Pratham 2009*)*

[11] A popular reading theory from the United States is the "simple view of reading". It states that comprehension is the product or sum of a reader's word decoding and listening comprehension skills. This theory has gained support over time (Kirby & Savage, 2008), so the components can be tested globally. Recent research, however, adds perceptual and memory variables. Accurate and rapid reading + instant access to word meanings + sufficient working memory = comprehension.

Comprehension happens only if input enters rapidly into working memory. (Of course instant language and content knowledge are needed; see below.) The brain areas related to comprehension and critical thinking are accessible *only* if the circuits related to perceptual learning are efficiently linked through practice to transfer electrochemical signals instantly. The simplest form is literal comprehension, which offers evidence that a message unit remained in working memory long enough for semantic connections to be made. Literal comprehension requires less processing than inferences. (See for example a clear discussion in McKoon & Ratcliff, 2017) and therefore early-grade tests should assess that essential literal comprehension step. **However, early focus on comprehension produces unreliable results**. Because comprehension is multidimensional, studies setting comprehension percentages as benchmarks to predict speed, have produced inconsistent results (e.g. Jukes et al., 2016). For example, EGRA comprehension tests mix literal and inferential questions, but difficulty levels vary across languages, and the mix is arbitrary. Some EGRA administrations, furthermore, allow students to look at the text while answering, and this practice contaminates the content of working memory that theoretically should be used to answer questions. To gauge the amount left in working memory, it is preferable, therefore, that questions are literal, rather than requiring inferences that take up extra time.

Early-grade reading in transparent orthographies (which represent 99% of the world's languages) should be easy, but has been subject to anglo-centric, complex, and common sense-based approaches. [12] For example USAID has been implementing a strategy based on early comprehension and has trained ministry staff and teachers on a large scale on incorrect principles. The result is very disappointing performance. In particular, failure is common among lower-income students. The culprits are whole-word instruction and a lack of practice. These issues are crucial for the first 2-3 grades, and delays carry into secondary school and beyond. Therefore, given delays in reading achievement, automaticity deserves monitoring in the early stages as well as beyond.

How to measure reading automaticity comparably?

Reading speed is the easiest aspect of reading to measure and has the greatest practical significance. Benchmarks for transparent orthographies can be developed from various studies that report grade-wise average words per minute (see for example Seymour et al., 2003 for European countries). Similarly, EGRA data exists for most countries of the world.

Norms and neurocognitive indicators suggest the following as prerequisite benchmarks:

- Forty-five to 60 words per minute should be the norm for all students by the end of grade 2, in just about every language and script.

[12] Stakeholders involved in education typically lack a background in cognitive science and often have limited knowledge about these neurocognitive issues. A feedback document on learning progressions for level 1 showed a focus on English and on middle-income country performance, which are inapplicable to other writing systems. Furthermore in 2017, 14 reading levels were developed, without any research evidence for that number.

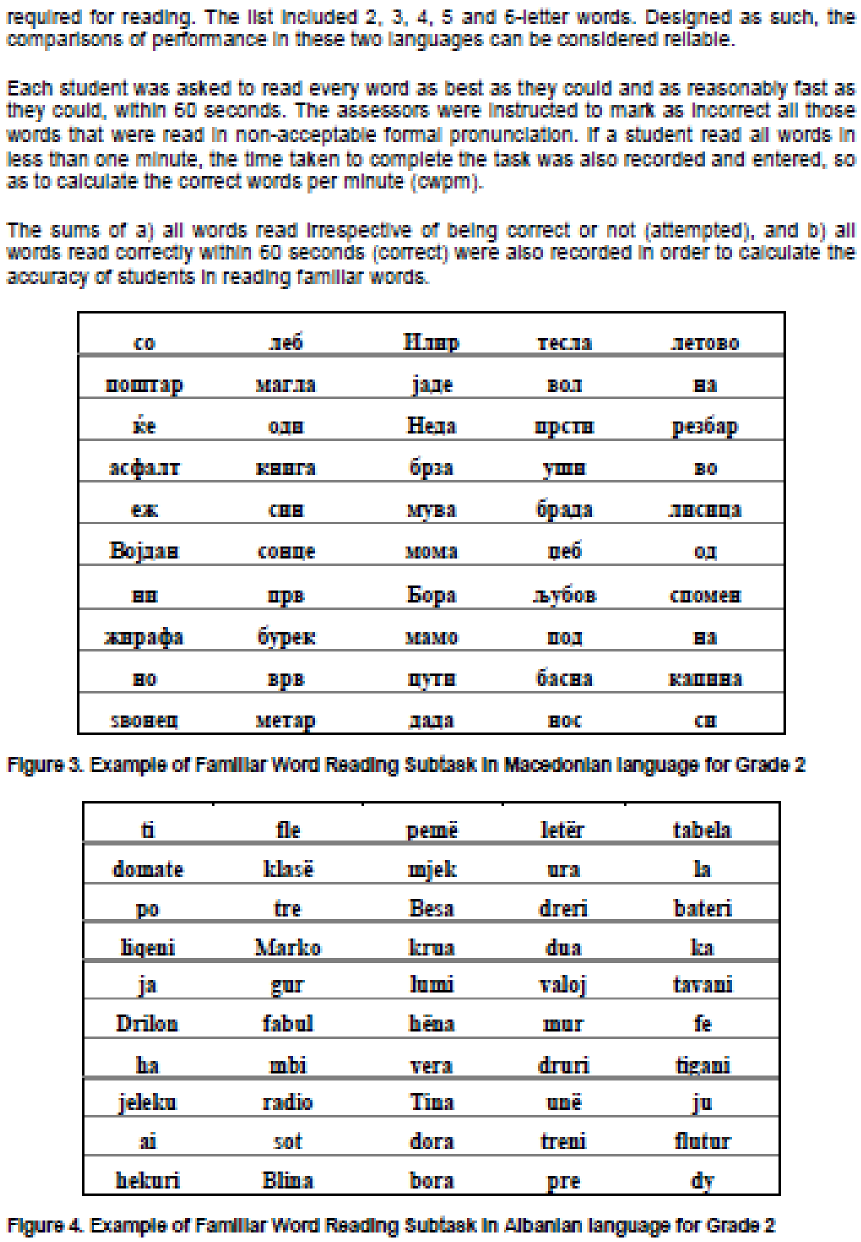
- By the end of grade 1 nearly all students should know nearly all letters of their script and the relevant decoding rules, or at least the characters prescribed for grade 1.[13]
- By the end of grade 2 at the latest students should read between 45 and 60 words per minute, given measurement methods for their languages.
- Students in grade 7 (1st secondary year) should read between 120 and 150 words per minute and show evidence of 75-80% comprehension at least.
  (120 wpm correspond to the 25%ile of US and Latin American data; Abadzi, 2013b).
- There should be zero students reading zero words per minute after grade 1 (special education aside).

The commonsense approach to reading has created a tendency to let stakeholders set their own reading benchmarks. But stakeholders often set speed benchmarks too low for comprehension. More research is needed on language equivalency methods, but the 45-60 wpm benchmark seems to reflect visual word form area activation in humans.

As mentioned elsewhere, it is feasible to **count words per minute** across languages and scripts. There are several methods, and also an average could be used (For methods see Abadzi, 2013). One easily understood comparison method builds a comparability bridge through reading lists of words that have similar numbers of letters in various languages (even in different scripts, although the perimetric complexity matters). For example, the Northern Macedonia EGRA (Step by Step Foundation) used 20 words of 2, 3, 4, 5 letters.

[13] For Chinese, and Japanese specific benchmarks exist; for syllabic scripts, the main matrices ought to be learned.

Figure 2.3: Word list reading in multiple languages



required for reading. The list included 2, 3, 4, 5 and 6-letter words. Designed as such, the comparisons of performance in these two languages can be considered reliable.

Each student was asked to read every word as best as they could and as reasonably fast as they could, within 60 seconds. The assessors were instructed to mark as incorrect all those words that were read in non-acceptable formal pronunciation. If a student read all words in less than one minute, the time taken to complete the task was also recorded and entered, so as to calculate the correct words per minute (cwpm).

The sums of a) all words read irrespective of being correct or not (attempted), and b) all words read correctly within 60 seconds (correct) were also recorded in order to calculate the accuracy of students in reading familiar words.

со	леб	Илир	тесла	летово
поштар	магла	јаде	вол	на
ќе	оди	Неда	прсти	резбар
асфалт	книга	брза	уши	во
еж	сви	мува	брада	лисица
Војдан	сонце	мома	цеб	од
ни	прв	Бора	љубов	спомен
жирафа	бурек	мамо	под	на
но	врв	цути	басна	капина
ѕвонец	метар	дада	нос	си

Figure 3. Example of Familiar Word Reading Subtask in Macedonian language for Grade 2

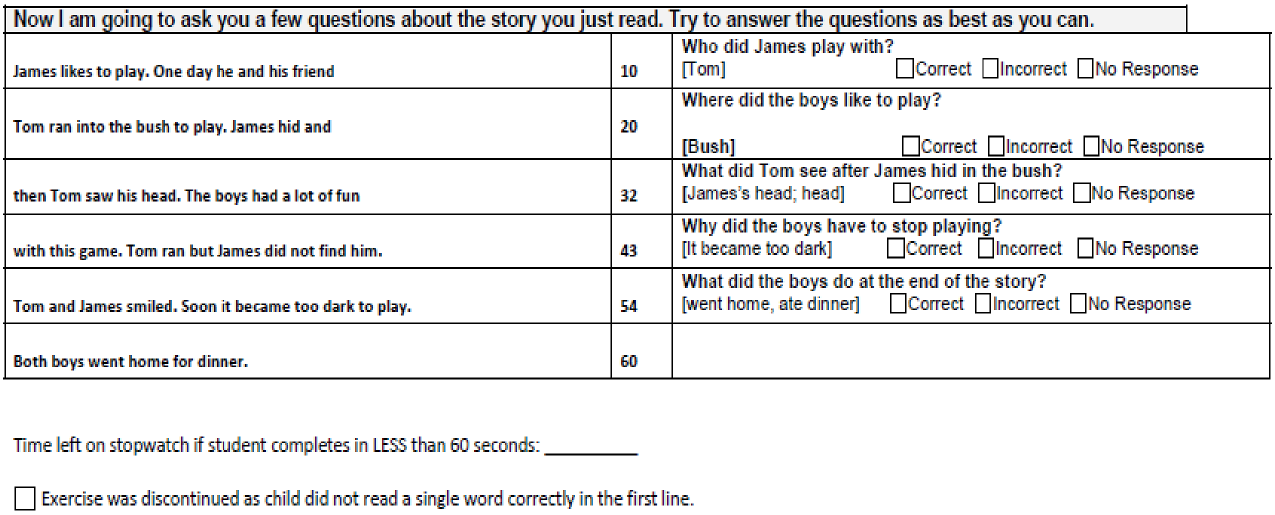
ti	fle	pemë	letër	tabela
domate	klasë	mjek	ura	la
po	tre	Besa	dreri	bateri
liqeni	Marko	krua	dua	ka
ja	gur	lumi	valoj	tavani
Drilon	fabul	hëna	mur	fe
ha	mbi	vera	druri	tigani
jeleku	radio	Tina	unë	ju
ai	sot	dora	treni	flutur
hekuri	Blina	bora	pre	dy

Figure 4. Example of Familiar Word Reading Subtask in Albanian language for Grade 2

For 'learning to read' (and learning to calculate), processing speed ought to be emphasized (i.e. items per minute). For 'reading to learn' (and calculating for math concepts), more knowledge-based tests could be used. This distinction could be made in the measurement strategy. This distinction offers advantages. "Reading to learn" requires test or item equating. But "learning to read" (or calculate) in grade 2 removes the need for content comparison across countries. The variable is speed and accuracy rather than responses to information items. Math items are, inevitably, very similar in various countries, but equivalency in reading items is really impossible. This may be one reason why colleagues hesitate to compare reading across languages and scripts.

Given the importance of automaticity, children in principle should be tested at the end of grade 1 in the reading curricula of their country. Only one or two EGRA subtests should be used, and time can be shortened to about 5 minutes. (A test battery of multiple basic skills can be created, to include math, oral language, and writing speed samples.) In principle, reading tests could be administered by computer and also involve silent reading. For oral reading, voice recognition has not yet been perfected, and it is unclear when this will become practical for large-scale use.

Figure 2.4: Example of a 60-word passage from EGRA (Early Grade Reading Assessment)



Now I am going to ask you a few questions about the story you just read. Try to answer the questions as best as you can.

James likes to play. One day he and his friend	10	Who did James play with? [Tom] □Correct □Incorrect □No Response
Tom ran into the bush to play. James hid and	20	Where did the boys like to play? [Bush] □Correct □Incorrect □No Response
then Tom saw his head. The boys had a lot of fun	32	What did Tom see after James hid in the bush? [James's head; head] □Correct □Incorrect □No Response
with this game. Tom ran but James did not find him.	43	Why did the boys have to stop playing? [It became too dark] □Correct □Incorrect □No Response
Tom and James smiled. Soon it became too dark to play.	54	What did the boys do at the end of the story? [went home, ate dinner] □Correct □Incorrect □No Response
Both boys went home for dinner.	60	

Time left on stopwatch if student completes in LESS than 60 seconds: ________

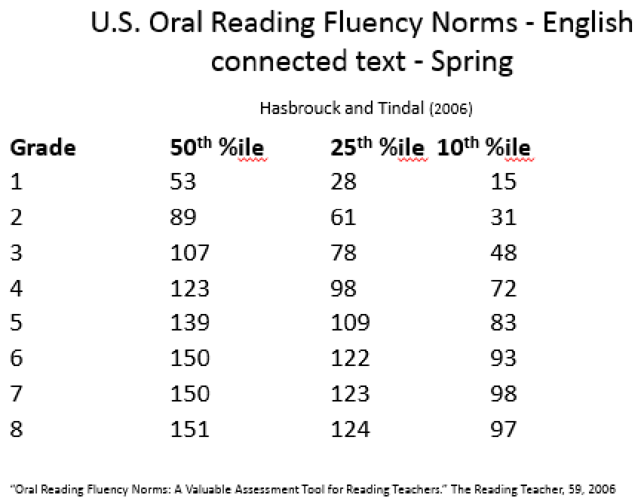
□ Exercise was discontinued as child did not read a single word correctly in the first line.

Because of the similar underlying processes, it is possible to start by using norms that exist, particularly in transparent orthographies and the Roman script. (e.g., English by Hasbrouck & Tindal, 2006 and various researchers for Spanish roughly coincide).

Do comprehension rates help estimate reading speed? This is difficult because the syntactic complexity of different languages and scripts may affect comprehension. The speed of comprehending individual words also changes. But longer and multiple samples of texts can be inputted into programs like google translate. Reading speed can be estimated with various methods, including taking multiple estimates. Then it may be possible to take averages.

Psychometricians among SDG stakeholders have raised concerns about equating reading test items across languages and scripts. But the neurocognitive approach mitigates this problem. Languages and script shapes are influenced by the common human DNA and are more similar than casually perceived. Therefore, the measurement of reading speed does not require using the same items. A text appropriate for grades 1-2 is sufficient (i.e. short sentences with common words). Remarkably, students do not even have to understand a language in order to read it fluently.

Table 2.1: Norms for English



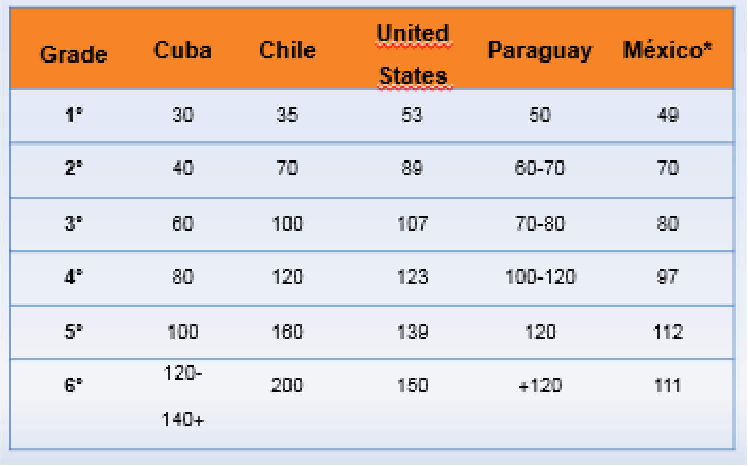
U.S. Oral Reading Fluency Norms - English connected text - Spring

Hasbrouck and Tindal (2006)

Grade	50th %ile	25th %ile	10th %ile
1	53	28	15
2	89	61	31
3	107	78	48
4	123	98	72
5	139	109	83
6	150	122	93
7	150	123	98
8	151	124	97

"Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers." The Reading Teacher, 59, 2006

Table 2.2: Norms of various countries in the Roman scripts tend to coincide[14] (50th percentile)



Grade	Cuba	Chile	United States	Paraguay	México*
1°	30	35	53	50	49
2°	40	70	89	60-70	70
3°	60	100	107	70-80	80
4°	80	120	123	100-120	97
5°	100	160	139	120	112
6°	120-140+	200	150	+120	111

Earlier GAML decisions on reading discounted the role of speed and the means to measure it across languages and scripts. The rationale presented was that research by the Research Triangle Institute showed that words per minute across languages cannot be compared. However, this is incorrect. One example is the Ethiopian EGRA, 2010.

[14] Cuba: Pérez Vilar, 1996, Chile: Reynols, 2005, Programa del Ministerio de Educación "Red Maestros –de-Maestros. Estados Unidos: Hasbrouck, J., y Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. Paraguay: Ministerio de Educación y Cultura de Paraguay, 2005. México: Tesis: Fluidez lectora por grado escolar en una muestra de niños mexicanos (2004) Distrito Federal. 1er Congreso Internac. de Innovación Educativa, México, UPN 2006. http://148.204.73.101:8008/jspui/bitstream/123456789/457/1/C10.doc. *Mexican data are from a simple of 16 schools (8 public and 8 private) of México City.

Testing by bypassing language and script differences. In principle, it is possible to bypass script complexity entirely and record brain voltage at the skull near the visual word form area. Electroencephalographic machines (EEG, event-related potentials) do this, and have been used often in reading experiments in the US. The process is time-consuming and rather expensive, but small samples worldwide can be taken for verification. (There are several comparative studies, e.g. Perfetti, Cao, and Booth, 2013). Cost and feasibility estimates for small international samples can be requested from the many scientists who do this work.

As mentioned earlier, some people excel in some skills over others. But schooling should enable everyone to reach a minimal competency in skills that predict performance of future tasks. What schools should produce, and the 4.1.1 target should monitor, is not only knowledge of single items and procedures but also performance that is fluent and overlearned (therefore relatively resistant to forgetting) in these core competencies.

Reading automaticity is a skill that is also at the heart of adult literacy monitoring (target 4.6). As with children, words per minute should be obtained. Adult illiterates have difficulty attaining automaticity, so attendance in literacy classes is no guarantee of fluency. In surveys, such as MICS, adults must be asked questions such as: “When you watch TV, do you read the words?” “If you glance at a newspaper, do you read the headlines?” “When you pass by stores do you read the signs?” If adults are asked to respond through tablets, latency may also be measured, that is the seconds that pass before a person starts a response.

Measuring language proficiency and comprehension (speed, accuracy)

Citizens of different countries have varied linguistic needs, and their ability to function in daily life should be assessed. Probably most countries in the world use multiple languages, often with religious implications. Citizens should be fluent in their own grammar and writing systems, as well as those used officially. The issues are beyond this document, but linguistic competencies could be dual or multiple.

Currently language is interwoven with reading, but the function originates in different brain regions and evolves differently. It is possible to read fluently without knowing a language; many countries are multilingual, and students may differ in their degree of official language command. For this reason, it is useful to consider the variables predicting comprehension and how to measure it. Even if a single test is given for reading and language, it may be possible to refine levels more precisely.

Despite apparent differences, languages have substantial similarities due to underlying information processes (Fedzedchkina, Chu, & Jaeger, 2017). Therefore, many aspects of world languages are comparable.

Language is learned through mechanisms well-honed by evolutionary pressures that predate humans; therefore normally developing brains in a community master the grammar and vocabulary essentials needed for daily communication in a community. However, students must know academic vocabulary, as well as background information, to understand text. Speakers of dialects must know the language of the textbooks, and in case of minority languages or official languages, the students must know the grammar and vocabulary applied in textbooks and in class. It may not be always the dialect they speak at home.

Language knowledge assessment is a topic for which much is written. Basic aspects are internationally used tests for the early grades, such as the Peabody Picture Vocabulary Test (PPVT),[15] the Woodcock-Johnson, and others. EGRA has a brief oral comprehension subtest and recently also a vocabulary subtest. These initially English-language tests tend to rely on noun names, such as body parts. This may be due to biases from English that has limited grammar. Vocabulary tests should also include adjectives, adverbs and verb endings, depending on the language. PASEC has been renewed with local-language content, and it may be viable for Francophone Africa.[16]

As with the other basic skills, speed of processing is crucial due to working memory limitations. For example, dialectical differences may delay comprehension and thus result in item characteristic curves that differ from the speakers of the formal version (Weber et al., 2015). Language proficiency and speed of access are important because they enable students who read slowly to predict likely alternatives to spellings that are unclear to them (in English or other languages). Thus, language proficiency could be a proxy for comprehension, along with reading speed.

Speaking fluency and comprehension speed (e.g. latencies) can be measured through various techniques used in scientific experiments, and it could be tried with small samples. Some techniques work with audio files of students. Oral testing, perhaps in computerized media, would also enable the assessment of students who are illiterate. Comprehension rates have not been used extensively in education, but the existing research makes this possible and potentially desirable. Technological approaches are also possible for comprehension tasks, such as the development of an automated sentence generator for the assessment of reading speed (Crossland et al., 2008).

Morphological awareness is particularly important for writing systems such as Arabic and Japanese (Muroya et al., 2017). But little consideration has been given in international comparisons of prefixes and suffixes. Such particles are language-specific. However,

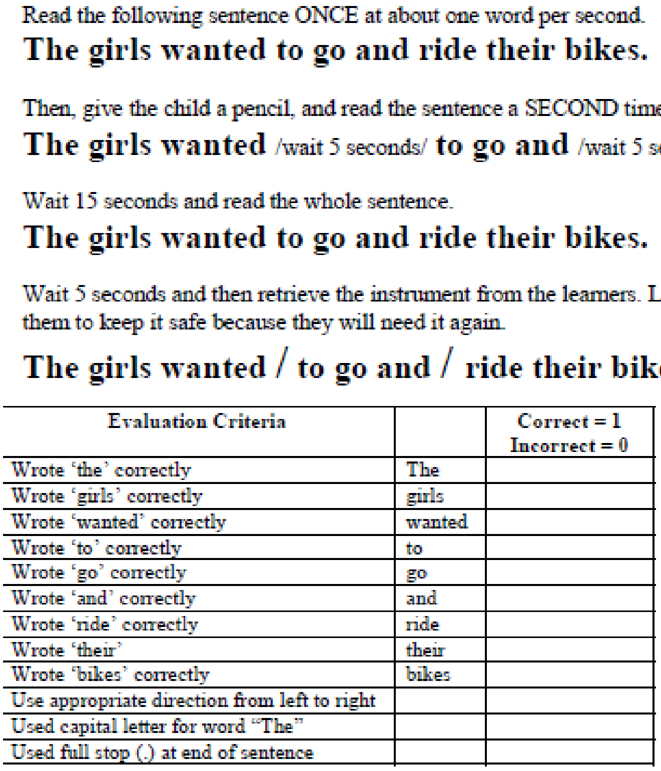
[15] PPVT-III assesses receptive vocabulary in adults and children older than 2 ½ years. On each item, participants are asked to point to the correct image from a panel of four images in response to a stimulus word. The interviewer records whether or not the participant selects the correct image. PPVT has a wide age range, easy administration, portability, and minimal amount of training required for testers. These features make the PPVT attractive in challenging field conditions and useful for longitudinal studies. However, it tests no grammar knowledge, which is essential in conjugated languages. Images represent only concrete items and cannot easily represent verb and noun conjugations.

[16] Often local languages use different spelling systems in the same country or in neighboring countries (e.g. Kichwa vs. Quechua; Limerick, 2018). The result is slower reading by those accustomed to a different spelling system. PASEC may have to deal with this issue.

morphological knowledge is likely to predict comprehension, so it may be worth devising ways to compare across languages. Frequency, length, location in words may be ways to compare.

Dictation tests are also appropriate for the lower grades. They focus on the ability to compose words from letters and spell correctly, although students may be writing too slowly to express thoughts.

Figure 2.5: Dictation test



Read the following sentence ONCE at about one word per second.

The girls wanted to go and ride their bikes.

Then, give the child a pencil, and read the sentence a SECOND time, grouping the words as follows:

The girls wanted /wait 5 seconds/ **to go and** /wait 5 seconds/ **ride their bikes.**

Wait 15 seconds and read the whole sentence.

The girls wanted to go and ride their bikes.

Wait 5 seconds and then retrieve the instrument from the learners. Leave the pencil with the child and tell them to keep it safe because they will need it again.

The girls wanted / to go and / ride their bikes.

Evaluation Criteria		Correct = 1
		Incorrect = 0

Wrote 'the' correctly	The	
Wrote 'girls' correctly	girls	
Wrote 'wanted' correctly	wanted	
Wrote 'to' correctly	to	
Wrote 'go' correctly	go	
Wrote 'and' correctly	and	
Wrote 'ride' correctly	ride	
Wrote 'their'	their	
Wrote 'bikes' correctly	bikes	
Use appropriate direction from left to right		
Used capital letter for word "The"		
Used full stop (.) at end of sentence		

Figure 2.6: Quick math problem solving and fast writing are needed

Measuring Writing Competencies for the Lower Grades

Writing fluency, correct spelling, connecting words in sentences that can be comprehended, are other basic skills. Writing fluency and % correct may prove to be a significant measure for the more advanced grades in lower-income countries. Research on writing speed and connectedness has been carried out in the US (Parker, Tindal, & Hasbrouck, 1991), and it offers some parameters that are potentially useful in all countries.

One early-grade writing test protocol is to give a theme, ask children to think for one minute and then write for 3 minutes. It is possible to measure numbers of written words, correctly spelled words, sensible and correct 2-word sequences, then calculate percentage of words legible, spelled correctly, sequenced correctly (Parker, Tindal & Hasbrouck, 1991). The following can serve as benchmarks.

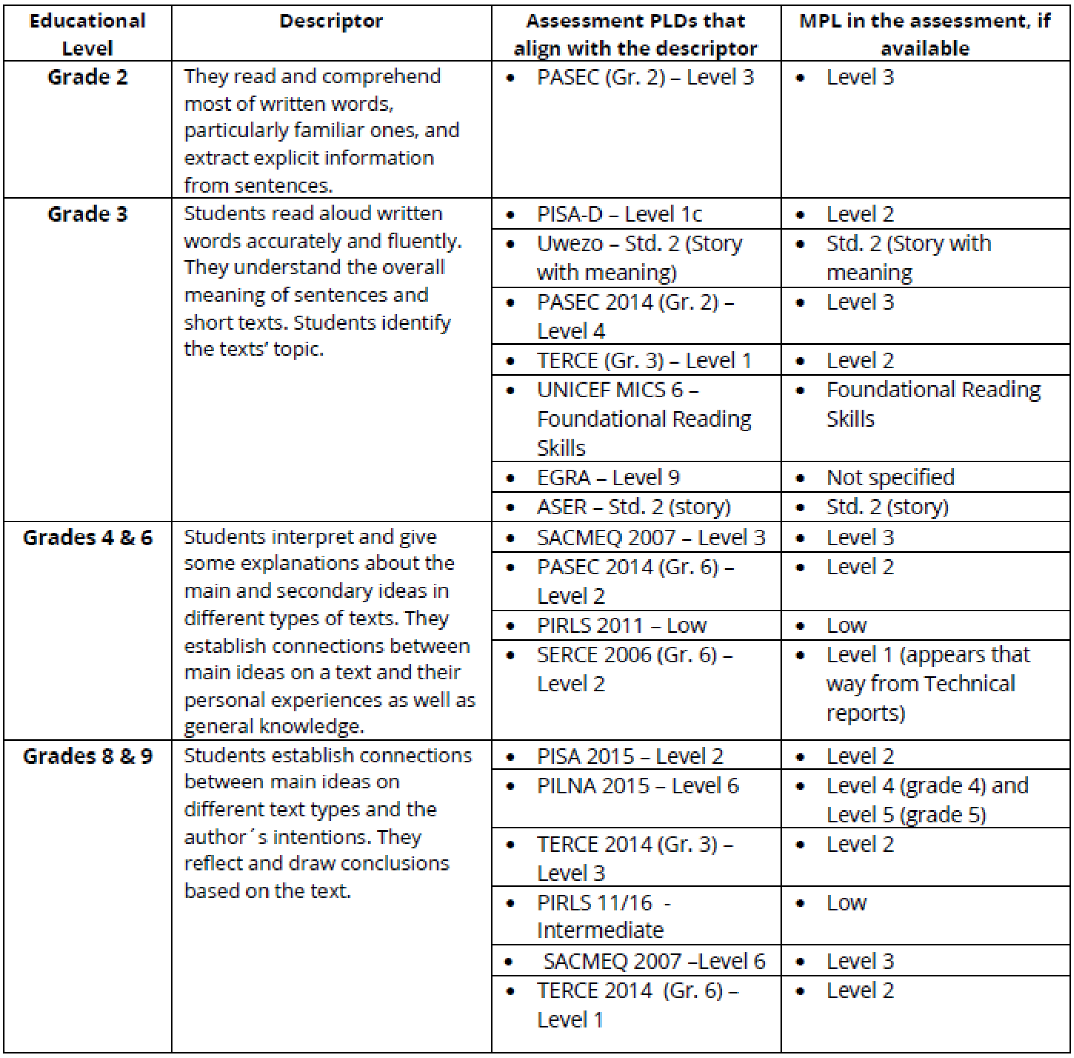
- Measures of writing length predicted various test scores in younger students (grade 2),
- measures of accuracy (% correct) predicted performance for 6th graders.
- Total words average were 25 in gr. 2, 39 in gr 4, 44 in gr. 6.

For **comprehension in higher grades, including secondary and post-secondary, a 1-minute test** developed in Portuguese can be adapted into other languages and scripts. It is a reading difficulties test, but the concept can be diversified. The **Teste de Idade Leitura** is a sentence comprehension test with a paper and pencil format, arranged in two A5 sheets, each with 18 sentences arranged in column (total of 36 sentences). At the end of each sentence, the last word is missing, and five words are presented in parenthesis. The participant is asked to select by circling which of the words correctly completes each sentence within a time limit of 5 min for as many sentences as possible. Rapid automatized naming was more associated with the 1-min test de idade leitura (across material, r=0.77, p<.001) than phonological awareness (Fernandez, Araújo, Sucena, et al., 2017). This Portuguese test can be adapted into other languages and scripts.

The 4th grade PIRLS assessment involves reading between 800 and 1000 word passages and answering multiple-choice and short-answer questions. The speed and language knowledge necessary for reading the text and answering the questions can be turned into benchmarks. For 4th grade and up, therefore, it is reasonable to use these specification and performance levels. For grade 2, clearly a different strategy is needed, as proposed.

As mentioned earlier, UIS has focused on measuring reading skills through existing regional or international tests, possibly with the addition of oral international tests (Table 2.3). However, these bypass fluency and focus on meaning and interpretation. Also some were developed for English or French, which have irregular spelling, so students are evaluated on whether they read familiar words. Furthermore some items allow guessing despite illiteracy (e.g., matching words to pictures). These may incorrectly assess reading, particularly in transparent orthographies. Slow readers who have, nevertheless, attained parallel processing, as well as students whose language knowledge is limited, are likely to be classified as illiterate.

Table 2.3: UIS proposal for minimum proficiency reading levels



Educational Level	Descriptor	Assessment PLDs that align with the descriptor	MPL in the assessment, if available
Grade 2	They read and comprehend most of written words, particularly familiar ones, and extract explicit information from sentences.	- PASEC (Gr. 2) – Level 3	- Level 3
Grade 3	Students read aloud written words accurately and fluently. They understand the overall meaning of sentences and short texts. Students identify the texts' topic.	- PISA-D – Level 1c	- Level 2
		- Uwezo – Std. 2 (Story with meaning)	- Std. 2 (Story with meaning
		- PASEC 2014 (Gr. 2) – Level 4	- Level 3
		- TERCE (Gr. 3) – Level 1	- Level 2
		- UNICEF MICS 6 – Foundational Reading Skills	- Foundational Reading Skills
		- EGRA – Level 9	- Not specified
		- ASER – Std. 2 (story)	- Std. 2 (story)
Grades 4 & 6	Students interpret and give some explanations about the main and secondary ideas in different types of texts. They establish connections between main ideas on a text and their personal experiences as well as general knowledge.	- SACMEQ 2007 – Level 3	- Level 3
		- PASEC 2014 (Gr. 6) – Level 2	- Level 2
		- PIRLS 2011 – Low	- Low
		- SERCE 2006 (Gr. 6) – Level 2	- Level 1 (appears that way from Technical reports)
Grades 8 & 9	Students establish connections between main ideas on different text types and the author´s intentions. They reflect and draw conclusions based on the text.	- PISA 2015 – Level 2	- Level 2
		- PILNA 2015 – Level 6	- Level 4 (grade 4) and Level 5 (grade 5)
		- TERCE 2014 (Gr. 3) – Level 3	- Level 2
		- PIRLS 11/16 - Intermediate	- Low
		- SACMEQ 2007 –Level 6	- Level 3
		- TERCE 2014 (Gr. 6) – Level 1	- Level 2

Note: Proposed minimum proficiency levels on the basis of content frameworks used in various international tests. Table in documents presented to the 5th GAML meeting, Hamburg, Germany, October 17-18, 2018.

Assessment of complex skills

One aspect of measurement involves the assessment of complex skills: These skills involve critical thinking, creativity, problem solving, communication and socio-emotional skills. There are assessments already in existence, such as the SimScientists program[17], which uses simulations to

[17] The SimScientists program in WestEd's Science, Technology, Engineering & Mathematics (STEM) program is comprised of a portfolio of research and development projects that focus on the roles that simulations can play in enriching science learning and assessment. The capabilities of technology allow modifications of simulation-based activities to offer accommodations for English learners and students with disabilities. Science simulations can be used in curriculum activities, as embedded, formative assessments, and as summative assessments. http://simscientist.org/home/index.php.

assess science learning (Vista, Kim, & Care 2018). Such assessments are challenging to develop and may be most suitable for high-income, high-performance environments.

Chapter 3. Math: A brain function for dealing with multiple objects in the environment

What constitutes "skills" in math and how can they be defined? This chapter has definitions and alternative measures.

The cognitive neuroscience of math

Unlike reading, numeracy is innate.[18] All beings are born to discriminate among concepts of quantity. Animals have a concrete number system that allows manipulation of a few items, but humans also do symbolic math, which they must learn (e.g., carrying borrowing, multiplication, division). Numeracy is related to brain connectivity, though some studies find a greater role than others for heredity. Clearly some people are much better at math than others, and better fed or more stimulated children tend to be better at math. But nearly all can refine their neurological system to solve daily 21st century problems.

Children are born with a number line that extends usually to the right of someone's visual field. The line is detailed for small numbers and logarithmic for larger ones, and it must be stretched through practice. Underlying this is an approximate and an exact numerical system. The approximate number system, also called "number sense," describes humans' and animals' ability to quickly size up the quantity of objects in their everyday environments.

The ability to use a mental number line appears to be dependent on a potentially inherent magnitude representation system. The inherent system results in estimations that conform to the natural logarithm (ln) of the number. In other words, the representations are compressed for larger magnitudes, such that the perceived distance between 2 and 3 is larger than the perceived distance between 89 and 90. Some children have deficits in the number-magnitude system, so their number line estimates might not conform to the natural log representation or might show less precision when making estimates based on this representation. Moreover, with schooling, children's number line estimates gradually conform to the linear mathematical system; the difference between two consecutive numbers is identical regardless of position on the number line (see Geary et al., 2007 for a review).

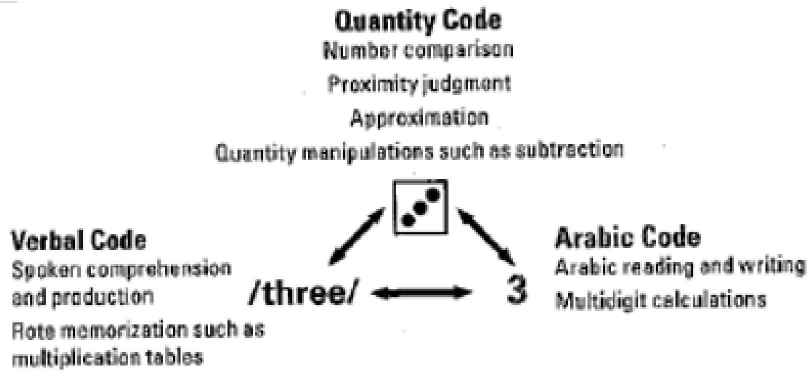
Math functions reside in multiple areas of the brain. One demonstration of this is the "triple code" of math: a sense of quantity, number name, and number symbol (Skagenholt et al., 2018). One big challenge is how to connect them into functional and quick calculation skills. Children can easily enunciate number words and link them to symbols, but they find it harder to link these to the sense of quantity. Practice and instruction must lead students to sense intuitively the quantity when words or numbers are evoked. These innate processes facilitate conceptual

[18] Left intraparietal sulcus activation during symbol processing task correlates with math fluency (D. Ansari research). One relevant OECD publication on the neuroscience of math is Looi, C. et al., 2016. The innate system becomes less important when children start to learn symbolic numbers (e.g. Geary & vanMarle in review).

understanding and high-level math. The usual focus on procedures in school, before students connect the magnitude part of the triple code to the others, may result in nonsense answers (5+7=57).

The brain has two numerical systems, exact and approximate. The latter signifies estimations, making quick guesses of about how many objects there are in a space. The accuracy of approximation, designated as a Weber fraction, correlates with academic performance (see the work of Ansari, Halberda, Feigenson, and others). For example, children who performed a dots game in a proper training fashion -- easiest to hardest -- scored much higher on the math test, getting about 80 percent of the answers correct, much higher than control students (Wang, Odic, Halberda, & Feigenson, 2016). Preschoolers who can make fine-grained discriminations of quantities have higher concurrent and later mathematics achievement than other children (in Geary et al., 2018; also Mou et al., 2016). Therefore, understanding cardinality (the quantity represented by number symbols) at age 4 gives children an early start in numerical relations and results in better math performance years later. This age benchmark and related activities could be a key skill to achieve in preschools of the world. (There are no gender differences in the US studies.) Children who do not understand the cardinality principle by age 4 are at risk of lifelong delayed math performance.[19]

Figure 3.1: The triple code of numeracy


Quantity Code
Number comparison
Proximity judgment
Approximation
Quantity manipulations such as subtraction
Verbal Code
Spoken comprehension and production
Rote memorization such as multiplication tables
/three/
3
Arabic Code
Arabic reading and writing
Multidigit calculations

Source: David DeSouza (ed), 2010. Mind, Brain, and Education, p. 187

Sensitivity to ratios also predicts math achievement. The brain has a ratio-processing system that is sensitive to magnitudes of non-symbolic ratios and may be ideally suited for supporting fraction

[19] The experimenter presents a pile of toys to a child and asks the child to give exactly 1, 2, 3, 4, 5, and 6 toys from the pile, beginning with 1 toy. If mistakes are made, the requested number is reduced. The "knower" criterion was 5-6. Knowers understand that 'one' refers to one and only one object of any kind, and "two knowers" understand the quantities represented by 'one' and 'two'. Eventually after 'three' or 'four', children understand that all number words refer to specific quantities and that each successive number word in their count list represents exactly one more than the word before it. The gap between number word and numeral knowledge closes substantially between 4- and 5-years and completely, at least for small values, by 6 years (Knudsen et al., 2015). While preschool children are learning the names and magnitudes of numerals, they are also beginning to accurately compare their relative magnitudes. (Geary & vanMehl, in submission).

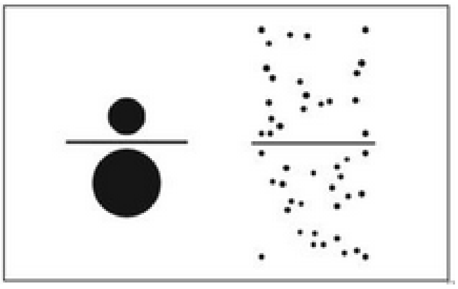
concepts. The effects were even evident at a university algebra entrance exam (Matthews et al., 2015).

Spatial reasoning and math

The skill of spatial reasoning is rarely assessed, but it is highly related to math. It is predictive of math as early as 6 months of age. It involves mental feats such as the ability to transform and rotate objects in "mental space." (Lauer & Lourenco, 2016). This skill can be tested through various tests. Visual form perception also predicts math performance (Jiaxin Cui et al., 2017).[20] For example, the quality of block play at four years of age was a predictor of high school mathematics achievement (Wolfgang et al., 2001). There is a relationship between young children's construction skills (such as playing with jigsaw puzzles and blocks) and strong number sense and success in solving mathematical word problems (Nath & Szücs, 2014). The link between spatial reasoning and math is so strong that it is "almost as if they are one and the same thing" (Dehaene, 1997, p. 125). It no longer makes sense to ask whether they are related (Chen and Mix, 2014). Spatial instruction will have a "two-for-one effect" that yields benefits in mathematics as well as the spatial domain (Verdine, Golinkoff, Hirsh-Pasek, & Newcombe, 2013, p. 13).

The practices of mathematicians also benefit from spatial reasoning. Many mathematicians stress that their work relies strongly on visual and spatial representations and forms of understanding (Whiteley, Sinclair, & Davis, 2015).[21] One implication may be a need for larger, more spread-out equations, as is the case for initial reading.

Figure 3.2: Ratio sensitivity



Many studies exist on the neural substrates of educationally relevant cognitive skills (Evans et al., 2016). These neural substrates are better developed in more affluent children, who have sufficient zinc and iron stores, known to affect math scores (multiple studies). For example, the

[20] After controlling for gender, age, and five general cognitive processes in teenagers, (choice reaction time, visual tracing, mental rotation, spatial working memory, and non-verbal matrices reasoning), visual form perception uniquely contributed to numerosity comparison, digit comparison, and exact computation; but it had no significant relation with approximate computation or curriculum-based math achievement (Jiaxin Cui et al., 2017).

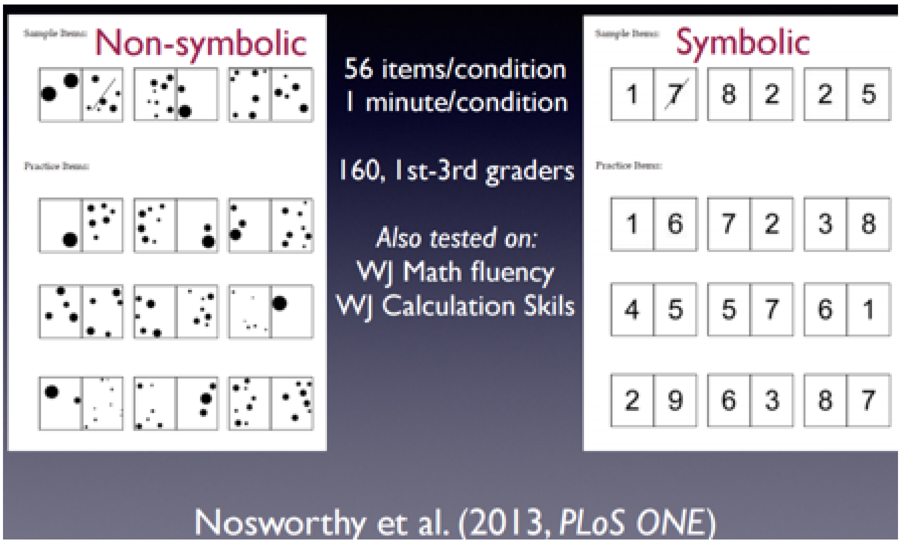
[21] https://www.kqed.org/mindshift/47301/five-compelling-reasons-to-teach-spatial-reasoning-to-young-children; Joan Moss, Catherine D. Bruce, Bev Caswell, Tara Flynn, and Zachary Hawes.

size and wiring of specific brain structures predicts how much an individual child would benefit from math tutoring (Supekar et al., 2013; Menon, 2014). Not surprisingly, children from more affluent families generally enter school with much greater math knowledge. It is crucial to measure and improve the numerical understanding of the poor (Siegler, 2009).

To ward off early delay and facilitate development, the sense of numerical magnitude should be strengthened early, i.e. in preschool (research by D. Ansari and others). Basic numerical magnitude processing, particularly during early development, constitutes an important scaffold for future skills. It is related to arithmetic skills, particularly symbolic skills and intuitive-cultural mapping. Children who become fluent in the relevant concepts later can group numbers and modify sets. Thus, strengthening numerical magnitude processing early is important for subsequent learning in math (see work by Ansari). Also understanding the relationship between size and number is critical for the development of higher math abilities. By combining number and size (e.g., area, density and perimeter), people can make faster and more efficient decisions. Students must progress from number sense to a sense of magnitude (Leibovitch et al., 2016).

The exact numerical system can be tested through mental math, oral or written. Children could be given single-digit exercises to fill in, but some testing must be done to see how long they take to write single digits. Accuracy of estimations can be tested through a 3-minute test (Nosworthy, 2013; figure 3.4). This has been tried in several countries that included the Gambia.

Figure 3.3: Magnitude estimations



56 items/condition
1 minute/condition

160, 1st-3rd graders

Also tested on:
WJ Math fluency
WJ Calculation Skils

Nosworthy et al. (2013, *PLoS ONE*)

The genetic system is insufficient for large quantities and complex manipulation. As with reading, students of course should demonstrate fluency in operations, including place value and borrowing of digits. They must demonstrate fluency in the spatial relations and calculations of fractions and mixed numbers (e.g. hours and minutes.)[22] These competencies facilitate performance in algebra and higher math. To understand daily math and the traps of advertisements, they ought to show understanding of essential statistics and probability (a difficult topic that is not innate). They must show fluency in conversions from various units to others that are needed in daily life (calculate time spans, convert between time units, calculate rates involving time). They must calculate percentages and have an intuitive sense of their magnitude. They must acquire ability to convert verbal problems into mathematical formulas and extract answers, particularly when problems have multiple stages. To do this, they must instantly read and understand the language.

Beyond these basics, multiple competencies unfold. Students must also be able to communicate results, represent them on paper, explain how they arrived at their solutions, find alternative solutions, and present arguments about them. They must also learn generalization, the ability to detect patterns and commonalities across different contexts. Developing increasingly abstract conceptions of those patterns is a key feature of mathematical proficiency. Students must learn to build mathematical rules that summarize various phenomena. Math use to solve problems across an increasing range of different contexts is an indicator of mathematical growth. Furthermore, the age at which children have the conceptual insight that number words represent specific quantities (cardinal value) is strongly related to their later number-system knowledge and is more consistently related to broader mathematics than to reading achievement, controlling for intelligence, executive function, and parental education levels. The age at which children acquire cardinal value knowledge, therefore, is central to later mathematical development and school readiness (Geary et al, 2017).

Substantial longitudinal relations exist between children's early mathematics achievement and their mathematics achievement much later. For example: math skill in 1st grade is linked to 7th grade performance, jobs and wages in the US (Geary et al., 2013). Knowledge of fractions and long division predicted math success in Canada (Siegler et al., 2012). A study of 14-year-olds in the US found that those who did well on a test that measured their "number sense" were much more likely to have achieved good grades in math classes. (Feigenson, Libertus, & Halberda, 2013).

The research presented earlier on math and reading points to some conceptual dilemmas. To read, students require formal or informal instruction; but in math it is possible to learn more complex calculations or hone estimations and the number line merely by interacting with the environment, e.g. selling at the market. Thus in math there are predictor variables that do not require formal instruction.

[22] In grade 2, number line estimation correlated significantly with calculation fluency (r = -0.27, $p < 0.05$) and math problem-solving (r = -0.52, $p < 0.01$). In grade 4, number line estimation correlated significantly with math problem-solving (r = -0.38, $p < 0.01$), but not with calculation fluency (Zhu et al., 2017). Subitizing predicted 18% of individual variance in Cubans over 12 years (2004-2016); R was 0.35 (Reigosa-Crespo, V., Torres, R., Mosquera, R., et al (in preparation).

Should the SDG-related tests look for evidence of instruction or is evidence of math accomplishment sufficient? Some students may not progress to higher grades where procedures are taught. If so, it is possible to develop a battery predictive of success whenever students actually study more advanced math. It seems, however, that GAML should monitor variables that show evidence of instruction rather than mere potential.

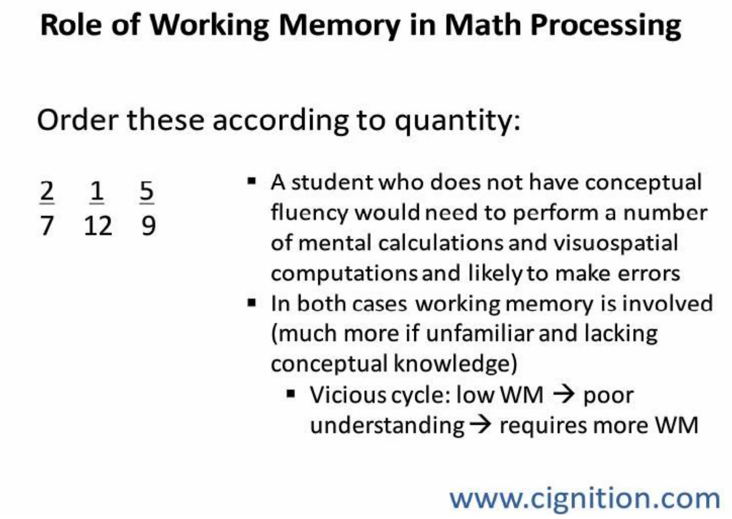
Proficiency and the importance of speedy retrieval in math

As mentioned earlier, working memory is important in numeracy. Approximately 60% of the variance in mathematics achievement may be accounted for by memory-related variables, such as working memory (Bailey et al., 2014). Speed and effortlessness is necessary. In performing simple mathematical processes (e.g., remembering basic facts) processing speed is necessary, so that working memory is freed up for more complex processes. Thus processing speed should be an essential component in math international comparisons. In the early grades, content is perhaps closer to the innate numeracy skills, so speed can be easily introduced based on students' knowledge of numerals and relations among them. (These in turn may be based on inherent number sense.) And accurate rapid calculations in the lower grades are essential for higher-level math. Practice and maturation develop approaches from immature (e.g., counting to solve addition problems) to more mature (e.g., remembering answers) problem solving strategies that in turn are associated with documented changes in the brain systems supporting these competencies (Qin et al., 2014).

There is much debate regarding the "drill and kill" style versus more conceptual, problem-solving based pedagogy.[23] Both types of training are necessary. A neuroscientific study of 8-9 year olds in the US showed that memorization of facts helps students move from counting to retrieval from memory and to solve complex calculations (Qin et al., 2014). Single-digit arithmetic calculation predicts performance on more complex math skills, illustrating the critical role that arithmetic fluency plays in building mathematical proficiency among students. (Price, Mazzocco, & Ansari, 2013). Thus, multiplication tables and additions must be memorized (probably in the school language). Slow and hesitant calculations and insensitivity to ratios would use up working memory and make an item, such as the one below, hard to solve.

[23] https://educhatter.wordpress.com/tag/math-wars/

Figure 3.4: Role of working memory in math processing



Role of Working Memory in Math Processing

Order these according to quantity:

$\frac{2}{7}$ $\frac{1}{12}$ $\frac{5}{9}$

- A student who does not have conceptual fluency would need to perform a number of mental calculations and visuospatial computations and likely to make errors
- In both cases working memory is involved (much more if unfamiliar and lacking conceptual knowledge)
  - Vicious cycle: low WM → poor understanding → requires more WM

www.cignition.com

In math, working memory can be depleted in various ways. Many students in the world do math in a language that is not their native language. Thus teaching math in a foreign language has cognitive costs. It involves storage related to various languages, and students need a longer time to solve the problem (Grabner et al., 2012). Given working memory capacity, this would carry a penalty in test scores. For low-income students of minority languages, this represents an additional burden.

The progression of math skills in school seems to have broad agreement in curricular documents. By the end of the third grade, students should be proficient in adding and subtracting whole numbers. Two years later, they should be proficient in multiplying and dividing them. By the end of the sixth grade, students should have mastered the multiplication and division of fractions and decimals (Math National Reading Panel of the US). This agreement facilitates testing at about the same grade levels.

How to measure early-grade numeracy?

The research suggests that four or five math-related features can be tested in the lower grades, and probably in the higher ones. Very strong candidates appear to be: the number sets test number of simple problems solved in 60 seconds, such as: speeded additions, subtractions, single-digit multiplications, ratio sensitivity and spatial transformations. Students could count backwards, and show evidence of engaging the number sense (though a test takes longer to score); make reasonably accurate numerical estimations. Furthermore, formal procedures taught in school must be tested. In principle, a battery of the non-redundant ones could maximize efficiency.

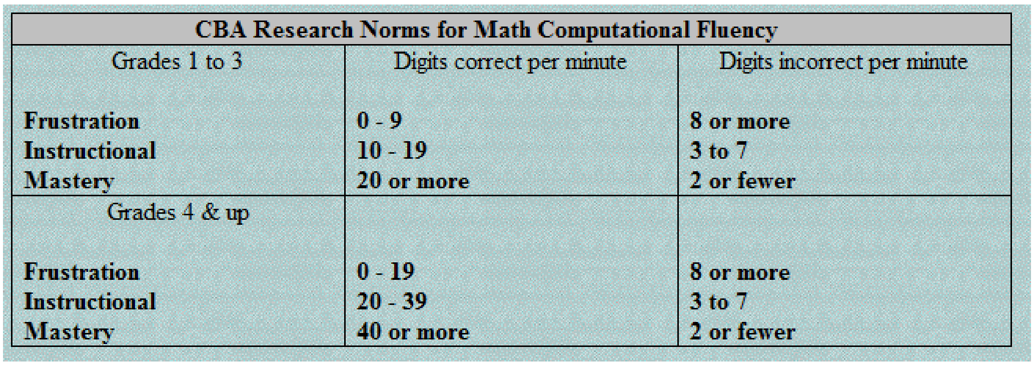
The most likely combination of tasks with the highest predictive validity of long-term outcomes would involve: (a) Fluency in solving basic addition and subtraction problems; e.g., how many items can be solved correctly in 60 seconds. (b) Understanding of the relative quantity of numerals, e.g., which is larger 42 or 29; for first graders values should be smaller; (c) Fluency of accessing quantities represented by numerals. This could be the Number Sets Test or a timed test to identify the larger of two small numerals e.g., 1 vs. 3, or 2 vs. 9. All of these will be correlated with one another and composites of them should be predictive of long-term outcomes in math, controlling other factors (e.g., Geary et al., 2013).

Studies of correct digits per minute are several. They are often referred to as Curriculum Based Assessments in the US (e.g. research by Fuchs and Fuchs). US research has consistently shown that student outcomes such as task completion, task comprehension, time on task, and learning growth significantly increase when the task demands represent an instructional level for both math and reading (Burns, 2002; 2007; Gickling & Armstrong, 1978, Gickling, Shane & Croskery, 1989; Treptow, Burns & McComas, 2006; VanDerHeyden & Burns, 2005 in Fuchs et al. 1998). Research done in the US has resulted in figures that could be used as benchmarks, for example: 14-31 digits correct per minute in grades 2-3 and 24-49 digits correct per minute in grades 4-5 (Burns, VanderHeyden and Jiban, 2006; citations in Fuchs, Fuchs, Hamlett, & Karns, 1998).[24] These proposed benchmarks require further validation and research.

Accuracy is at least as important as speed. For the US 70-85% correct completion was proposed (Glicking and Thompson, 1985; in Fuchs et al. 1998).

In principle, numeracy essentials could be tested orally in grade 1 and also grades 2 and 3. Differences in average performance could be compared to existing norms. These largely come from higher-income countries, but math performance is required in life. Lower income students must somehow rise to the demanded level.

Table 3.1: Norms on early grade math speed, indicating processing fluency and automaticity (Fuchs, Fuchs, Hamlett, & Karns, 1998)



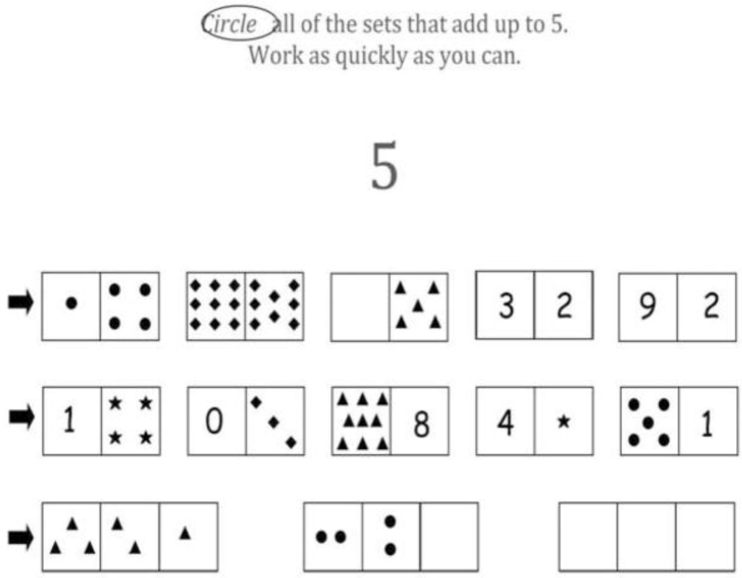
CBA Research Norms for Math Computational Fluency		
Grades 1 to 3	Digits correct per minute	Digits incorrect per minute
Frustration	**0 - 9**	**8 or more**
Instructional	**10 - 19**	**3 to 7**
Mastery	**20 or more**	**2 or fewer**
Grades 4 & up		
Frustration	**0 - 19**	**8 or more**
Instructional	**20 - 39**	**3 to 7**
Mastery	**40 or more**	**2 or fewer**

The **Number Sets Test** is a group-administered pencil-and-paper measure of the speed and accuracy with which children can identify the number and quantity of sets of objects and combine these with quantities, as represented by Arabic numerals. The combination thus potentially taps

[24] References to these articles are in Fuchs et al., 1998.

into critical features of number sense (Geary et al., 2007). Additional quantitative measures create a score (Geary et al., 2018) and include the percentage of addition facts correctly retrieved from long-term memory, the detection of double counting errors on a counting knowledge task, and the accuracy (i.e., degree of error) of the child's placement on a number line task. Performance on each of these measures is predictive of later mathematics achievement and is influenced by schooling (e.g., Geary, Bow-Thomas, & Yao, 1992; Jordan et al., 2003; Siegler & Booth, 2004). The measures may also capture components of children's early number sense.[25] As an example, performance on number line task is dependent on the same area of the parietal cortex that supports processing of magnitude and general quantity (Zorzi, Priftis, & Umiltá, 2002, in Geary et al, 2009). The test is timed with a stopwatch, and children get 60 or 90 seconds per page. It includes subtests: number sets, number estimation, counting knowledge, and addition strategy assessment. It has been used for kindergartens, but could be adapted for the early grades with larger numbers and fewer game-like features. For example, one task is to determine as quickly and accurately as possible if pairs or trios of object sets, Arabic numerals, or a combination of these match a target number (5 and 9; Figure 16). To diagnose children with mathematics learning disability, cutoffs have been developed that range from lenient (< 30th percentile) to restrictive (< 5th or 10th percentile). These could be used as benchmarks. The sensitivity measure from the Number Sets Test at the beginning of first grade is strongly predictive of standardized mathematics achievement at the end of first grade, controlling for intelligence and other factors (Geary, 2011), and predicts employment-relevant quantitative skills more than six years later (Geary et al., 2013).

Figure 3.5: Number sets test sample

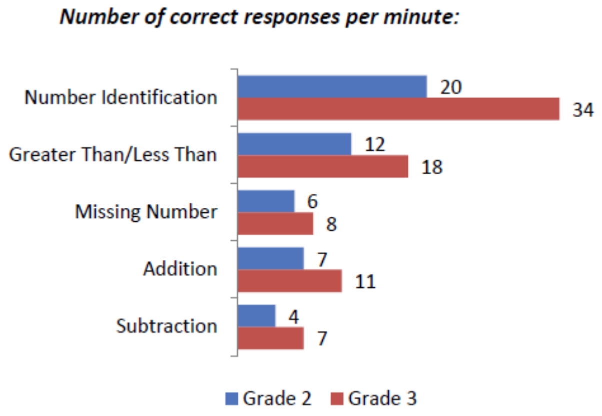

Circle all of the sets that add up to 5.
Work as quickly as you can.
5
3
2
9
2
1
0
8
4
1

[25] The nonsymbolic items in the Number Sets Test may engage students' intuitive number sense and their fluency in integrating this with their emerging knowledge of the quantities represented by numerals (D. Geary, personal communication).

The **Early Grade Math Assessment** (EGMA; Research Triangle Institute) has been administered extensively and in many countries. Attainment rates in subtests can certainly be of use. However, EGMA does not assess calculation speed. Students are asked to read numbers per minute, but this engages only the symbol-sound part of the triple code. (In effect it constitutes reading.)[26]

Several other tests have been used internationally and norms exist. Examples are Peabody individual achievement test math and the Woodcock Johnson math fluency test. PASEC and SACMEQ include math tests for the lower grades. It is unclear how low-literacy students deal with them, but TIMSS scores in the Arab world, where students read slowly, seem to be affected by reading speed.[27]

Figure 3.6: USAID usage of fluency in EGMA, though the expected number of items per minute number is too high (RTI, Dukkala Abda, Morocco)


Number of correct responses per minute:

Students who do not have automatic ability on simple addition and subtraction problems (e.g., 4+2 or 9-5) will likely have increasing difficulty with more complicated numbers and operations.

Note that 2nd and 3rd graders in the U.S. are expected to do 100 basic addition and subtraction operations per minute.

Citizen-administered tests can certainly be used for reading as well as for math. The ASER test (as well as similar ones in Kenya and Uganda) request simple calculations, which ought to be learning in grades 1-2. Estimation or spatial reasoning items could be added. Tests are typically applied at the household level to children based on the last grade they completed in school, but children are often of widely varying ages, and some may not even know their age.

One test resembles the citizen-administered tests is UNICEF's Multimode Cluster surveys MICS). The surveys (now in MICS 6) have considerable appeal. Over 40 countries have signed up for it, and include countries such as North Korea and Sierra Leone, which have limited comparative data. UNICEF works with national statistical officers and therefore has support.

[26] For early grades in developing countries, an assessment of numeral naming (ensures they know enough to do the tests), numeral comparison (more than/less than, gets at gets at conceptual understanding of magnitudes of numerals), number sets (fluency in accessing quantities associated with numerals), and speeded addition, subtraction (gets on fluency in foundational arithmetic) would be a quick to administer a screener (D. Geary, personal communication).

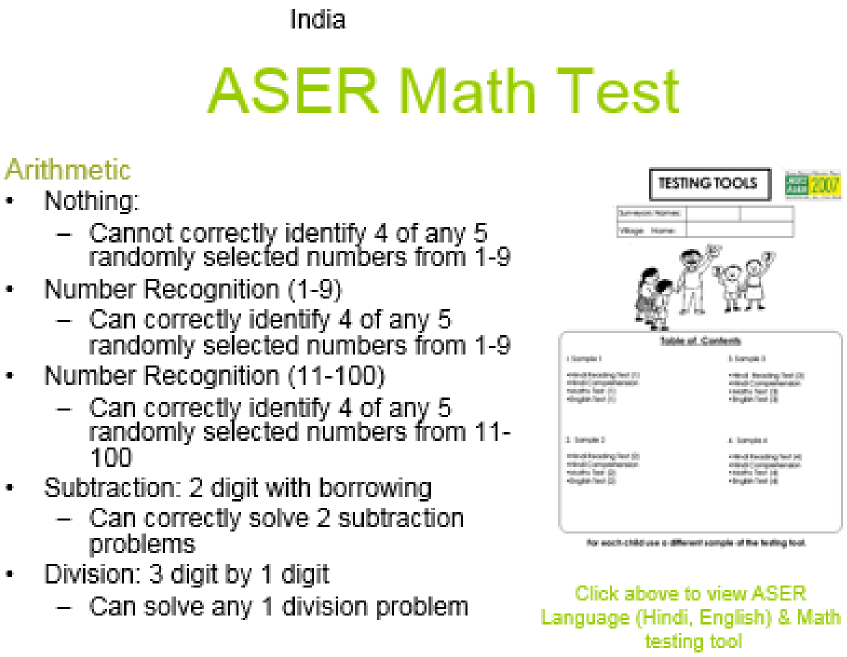
[27] Reading research suggests that numbers and equations should be in large and spaced fonts.

More systematic assessments can be done as part of UNICEF's MICS education module or other household surveys. However, MICs do not use timing explicitly. Though people cannot easily define automaticity and fluency, they sense time limits. Thus, MICS and ASER use implicit timing; that is, the interviewer decides how long to wait for an answer, perhaps 5-10 seconds. The wait time is indicative of people's usual processing speed. This implicit timing could be measured and used as a rough indicator of items calculated per minute. Thus, it may be possible to convert the implicit timing to a 'words per minute' estimate. Obviously comparative studies are necessary to establish the reliability of this implicit measure.

ASER and Uwezo will be used for SDG monitoring. Reading fluency is reported approximately, at the paragraph, sentence, word and letter level. Research by Pratham created some rough equivalences of 'words per minute', but these would have to be reassessed with a group of children. By using the citizen-led assessments, out of school children would also be reached. One challenge is how to convert these test scores to concepts resembling other tests.

Citizen tests are also offered at post-primary levels, to monitor the skills of out-of-school populations. In 2017, ASER focused on youth aged 14-18 years old who have moved just beyond primary school age. The report explored a wider set of domains and compared tasks that are simple to administer and easy to understand. The ASER test for youth (2018) was conducted with the participation of local partner organizations.

Figure 3.7: ASER math



India

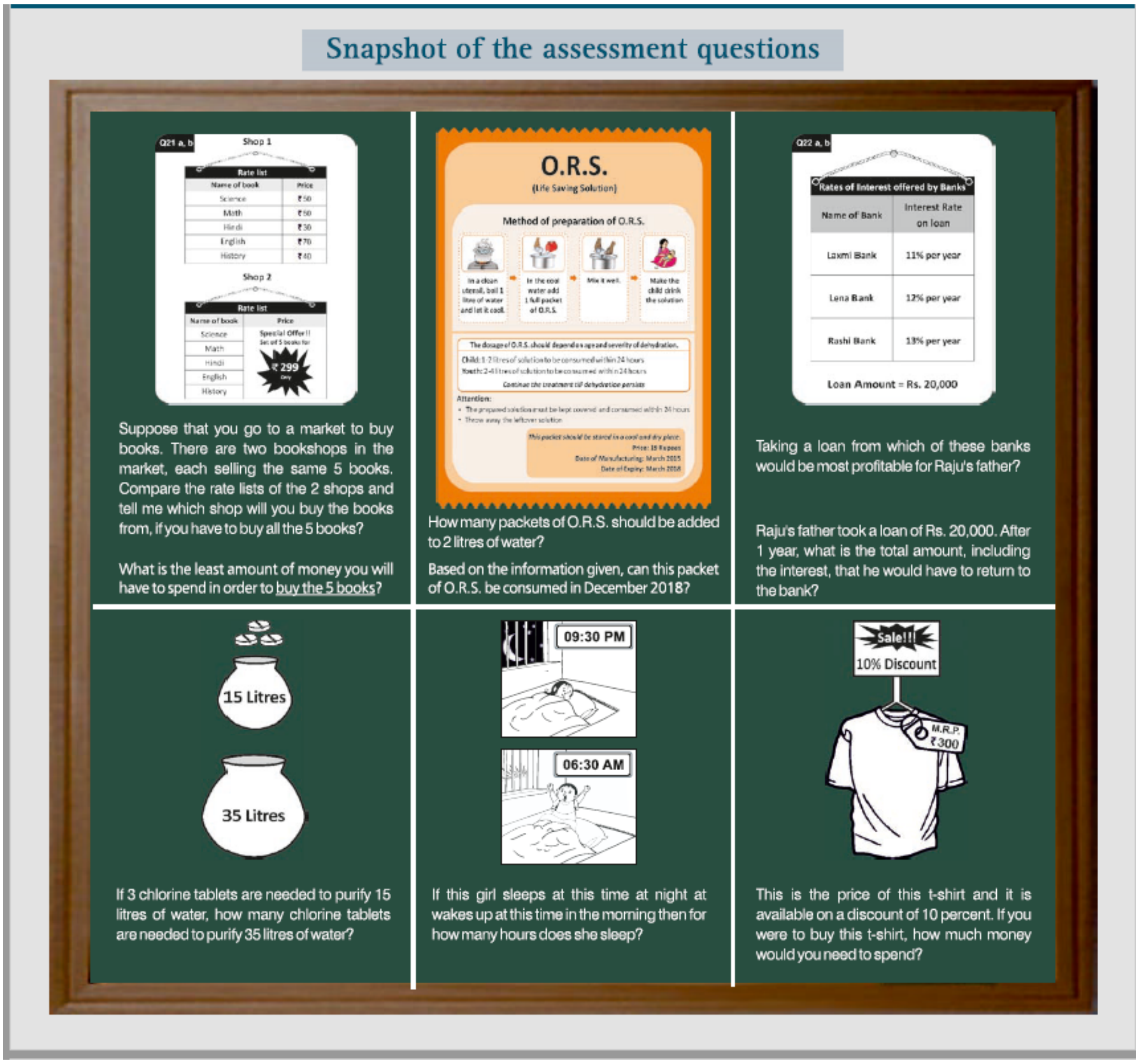
ASER Math Test

Arithmetic
- Nothing:
  - Cannot correctly identify 4 of any 5 randomly selected numbers from 1-9
- Number Recognition (1-9)
  - Can correctly identify 4 of any 5 randomly selected numbers from 1-9
- Number Recognition (11-100)
  - Can correctly identify 4 of any 5 randomly selected numbers from 11-100
- Subtraction: 2 digit with borrowing
  - Can correctly solve 2 subtraction problems
- Division: 3 digit by 1 digit
  - Can solve any 1 division problem

Click above to view ASER Language (Hindi, English) & Math testing tool

The ASER for youth results were depicted as follows:

Figure 3.8: ASER for youth sample



Snapshot of the assessment questions

Shop 1

Rate list	
Name of book	Price
Science	₹50
Math	₹50
Hindi	₹30
English	₹70
History	₹40

Shop 2

Rate list	
Name of book	Price
Science	Special Offer !! Set of 5 books for ₹ 299 Only
Math	
Hindi	
English	
History	

Suppose that you go to a market to buy books. There are two bookshops in the market, each selling the same 5 books. Compare the rate lists of the 2 shops and tell me which shop will you buy the books from, if you have to buy all the 5 books?

What is the least amount of money you will have to spend in order to buy the 5 books?

O.R.S.

(Life Saving Solution)

Method of preparation of O.R.S.

In a clean utensil, boil 1 litre of water and let it cool. → In the cool water add 1 full packet of O.R.S. → Mix it well. → Make the child drink the solution

The dosage of O.R.S. should depend on age and severity of dehydration.

Child: 1-2 litres of solution to be consumed within 24 hours

Youth: 2-4 litres of solution to be consumed within 24 hours

Continue the treatment till dehydration persists

Attention:

- The prepared solution must be kept covered and consumed within 24 hours
- Throw away the leftover solution

This packet should be stored in a cool and dry place.

Price: 15 Rupees

Date of Manufacturing: March 2015

Date of Expiry: March 2018

How many packets of O.R.S. should be added to 2 litres of water?

Based on the information given, can this packet of O.R.S. be consumed in December 2018?

Rates of Interest offered by Banks	
Name of Bank	Interest Rate on loan
Laxmi Bank	11% per year
Lena Bank	12% per year
Rashi Bank	13% per year

Loan Amount = Rs. 20,000

Taking a loan from which of these banks would be most profitable for Raju's father?

Raju's father took a loan of Rs. 20,000. After 1 year, what is the total amount, including the interest, that he would have to return to the bank?

If 3 chlorine tablets are needed to purify 15 litres of water, how many chlorine tablets are needed to purify 35 litres of water?

If this girl sleeps at this time at night at wakes up at this time in the morning then for how many hours does she sleep?

This is the price of this t-shirt and it is available on a discount of 10 percent. If you were to buy this t-shirt, how much money would you need to spend?

Figure 3.9: ASER for youth results

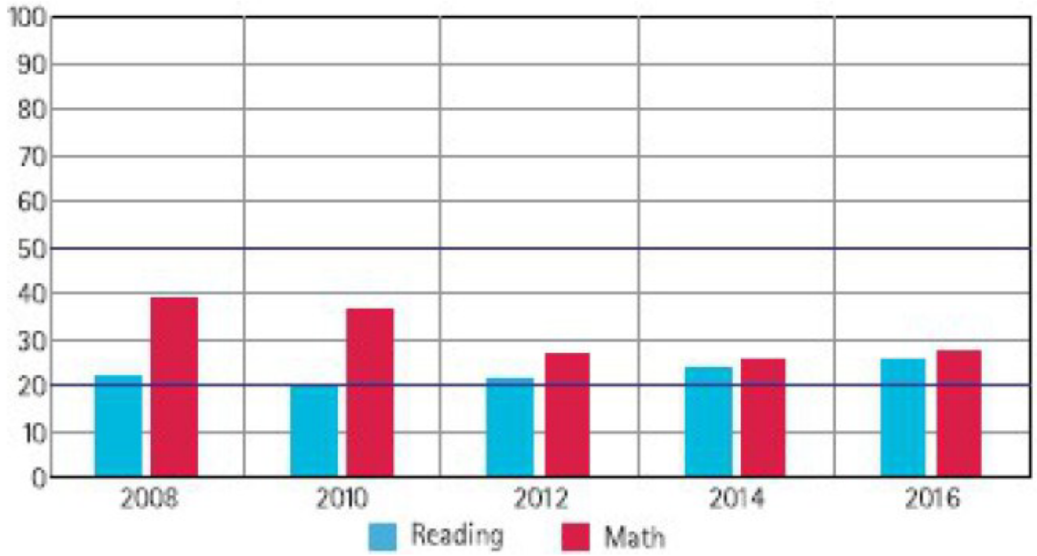


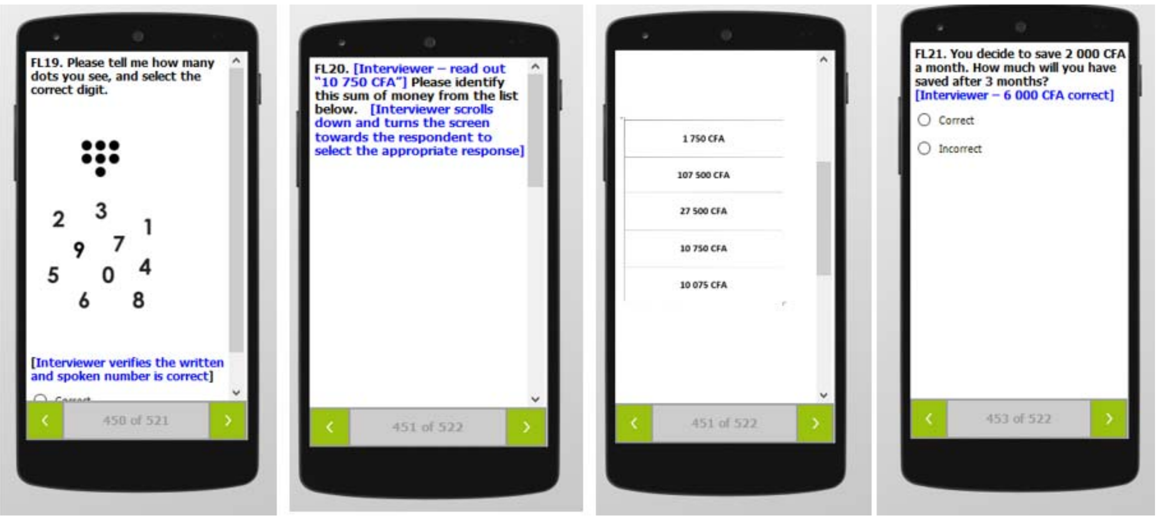
Figure 2: Children in class III who are at 'Grade Level' 2008-2016. The lack of commensurate ability with grade level is evident. Source: ASER 2016

Smartphone testing of numeracy for adults and youth

It is theoretically possible to test teenagers and adults at a distance, using cell phones. One NGO (My Oral Village, Brett Matthews) has collected data in Myanmar and Cote d'Ivoire using cellphone screens, as below. The data showed significant deficiencies in identifying the correct solutions. Reading speed and spacing could inhibit solutions, but also among the low-literates or illiterates, the number line may develop poorly for larger quantities. Ultimately many citizens could not use cellphone banking apps, like m-pesa.

The NGO "My Oral Village" has been developing indicators on the basis of this content (Matthews, personal communication, January 16, 2018). Test administrations instructions are as follows, used in Myanmar and Cote d'Ivoire

Figure 3.10: My Oral Village Math Testing for Illiterate Adults



FL19. Please tell me how many dots you see, and select the correct digit.

[Interviewer verifies the written and spoken number is correct]

○ Correct

FL20. [Interviewer – read out "10 750 CFA"] Please identify this sum of money from the list below. [Interviewer scrolls down and turns the screen towards the respondent to select the appropriate response]

1 750 CFA

107 500 CFA

27 500 CFA

10 750 CFA

10 075 CFA

FL21. You decide to save 2 000 CFA a month. How much will you have saved after 3 months? [Interviewer – 6 000 CFA correct]

○ Correct

○ Incorrect

Table 3.2: Questions asked and scoring

FL19	ALL	[Tablet displays a fixed number of dots (about 1 cm in diameter, all the same color) between 1-9 on the screen. Swipes to next page where the digits 1-9 are shown, along with the previous page dot configuration, small in the upper right corner.] Please tell me how many dots you see, and select the correct digit. [Interviewer verifies the written and spoken number is correct]	1 = Correct 2 = Incorrect
FL20	ALL	[Interviewer – read out "107500 kyat"] Please identify the sum of money I just read out from the list below. [Interviewer turns the screen towards the respondent to select the appropriate response]	1 = 17500 Ks 2 = 1075000 Ks 3 = 275000 Ks 4 = 107500 Ks 5 = 100750 Ks
FL21	ALL	You decide to save 20000 Ks a month. How much will you have saved after 3 months? [Interviewer – 60000 K correct]	1 = Correct 2 = Incorrect

Preschool-level competencies related to math performance in school

Many studies underline the importance of intervention at the preschool level. One study estimated the causal links between preschool mathematics learning and late elementary school mathematics achievement using variation in treatment assignment to an early mathematics intervention as an instrument for preschool mathematics change. Estimates indicate that a standard deviation of intervention-produced change at age 4 is associated with a 0.24-SD gain in achievement in late elementary school. This impact is approximately half the size of the association produced by correlational models relating later achievement to preschool math change. Implications for measurement and instruction are significant (Watts, Duncan, Clements, & Sarama, 2017).

In addition to the many programs discussed earlier, OECD has developed an early childhood learning measurement program to assess multiple dimensions of children's development that are highly predictive of future learning outcomes. They include various math and language variables as well as working memory, self control, and pro-social behaviors (Figure 20a). The study is expected to start collecting data in 2018 in about 18 countries.[28] Data will be collected from about 3000 students per country from 200 random sites. Each sub-component will take 15 minutes to administer on tablets. Some specifics were unclear in early 2018, but some components may be useful for repeated measurements to 2030.

The OECD child study measurement dimensions include:

- Emergent literacy skills: vocabulary, listening comprehension, phonological awareness
- Emergent numeracy skills: working with numbers, counting, shape and space, measurement and patterns,
- Self-regulation: working memory, mental flexibility, self-control,
- Social and emotional skills: trust, empathy, pro-social behaviors.

Potential neuroimaging assessments of early-grade numeracy

Like reading, math can be tested through event-related potentials. Quantitative measures of brain structure and intrinsic brain organization can provide a more sensitive marker of skill acquisition than behavioral measures (Supekar et al., 2013). In fact, EEG reveals signatures of executive control processes in arithmetic strategies and is a powerful tool for exploring between-group differences in arithmetic (Hinault & Lemaire, 2016). The cost and feasibility for small samples can be explored. Also, artificial intelligence programs may be useful for teaching math tailored to individual needs. On that basis, adaptive testing is possible, if students can be tested through computers. Some countries may be able to do this. In principle for such countries, e-testing could be tried in grade 2 as well.

Global Framework for Reference in Mathematics and Reading

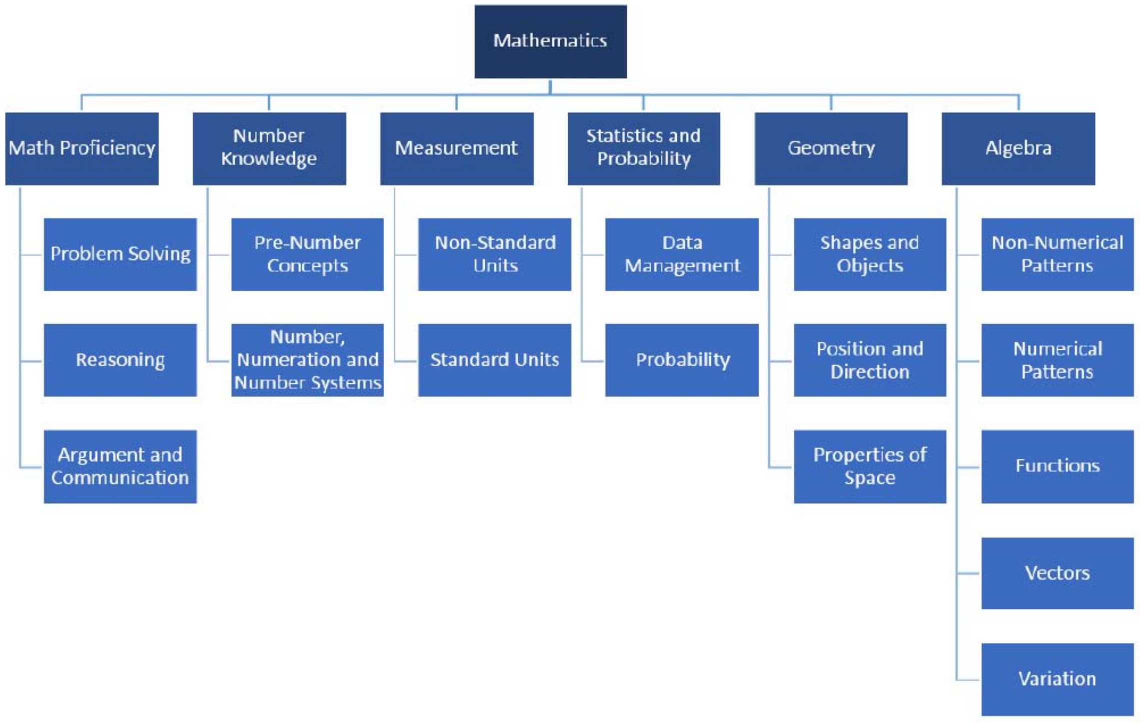
The UIS and UNESCO's International Bureau of Education (IBE) have developed draft global frameworks of reading and mathematics, which aim to help national and international stakeholders map and align curricula with national or international assessment frameworks. Taking into account the results of the global consultations, the final frameworks will soon be available as online references. The tools will allow users to automatically map their national or international assessments to the Global Framework of Reference by answering a series of questions.

[28] http://www.oecd.org/edu/school/international-early-learning-and-child-well-beingstudy.htm.

UIS and its consultants decided to define minimum proficiency levels in math according to the levels and specifications set in various international tests (Table 3.1). This is a practical means to define proficiency given existing data. Minimum proficiency should imply "functional numeracy". The working group on numeracy provided a tentative description of the proposed minimal proficiency level in numeracy as follows:

"The respondent is able to carry out basic mathematical processes in common, concrete contexts where the mathematical content is explicit, with either little or no text and minimal distractors. Tasks usually require simple one-step processes, and may involve understanding of representations of numerical entities (e.g., positions on a number line up to 100), performing basic arithmetic operations in reference to written or visual representations of quantities; understanding simple proportions (e.g., fractions or percentages such as 1/2 or 50%); locating, identifying, and using elements of simple graphical or spatial representations; and understanding basic information about everyday measurement systems such as regarding time, length or weight."

Figure 3.11: The mathematical domains that various curricular frameworks cover in principle



Note: Graphic in documents presented to the 5th GAML meeting, Hamburg, Germany, October 17-18, 2018.

Table 3.3: Minimum math assessment levels in various international measures

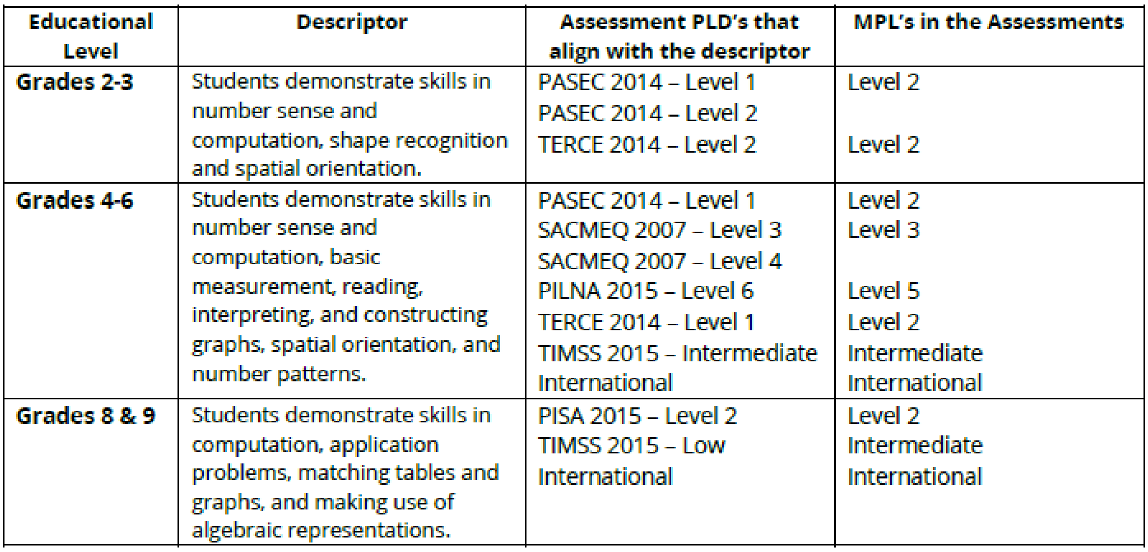
Schooling Level	Minimum Math Assessment Level
End of lower secondary	PISA level 2 TIMSS low international
End of Primary	SACMEQ level 3 / level 4 PASEC grade 6 level 1 PILNA level 6 TERCE grade 6 level 1 TIMSS 4 intermediate international benchmark
Grades 2/3	PASEC grade level 2 TERCE grade 3 level 2 Further work to align MICS/ASER/Uwezo

Note: S. Montoya. November 2018. GAML 5 exploring options on reporting 4.1

Currently no fluency measure exists for math; the emphasis is on knowledge and ability to carry out mathematical operations, however laboriously.

Unlike literacy, where the attainment of parallel processing gives a relatively clear range of words per minute to measure, numeracy does not have an easily identifiable automaticity indicator for math. Students continue to learn new concepts in various grades that must be automatized. The closest existing practical indicator for fluency comes from the Fuchs et al. study, which defines 9-19 single-digit operations per minute as "instructional" level. Such a test is brief and could be included in some measures. For example, in PASEC and TERCE grades 2 and 3, students are sensibly measured on number sense and computation, but not on fluency. Also attainment of the cardinality principle could be measured, in subtests borrowed from the Number Sets Test.

Table 3.4: Minimum proficiency levels for mathematics



Educational Level	Descriptor	Assessment PLD's that align with the descriptor	MPL's in the Assessments
Grades 2-3	Students demonstrate skills in number sense and computation, shape recognition and spatial orientation.	PASEC 2014 – Level 1 PASEC 2014 – Level 2 TERCE 2014 – Level 2	Level 2 Level 2
Grades 4-6	Students demonstrate skills in number sense and computation, basic measurement, reading, interpreting, and constructing graphs, spatial orientation, and number patterns.	PASEC 2014 – Level 1 SACMEQ 2007 – Level 3 SACMEQ 2007 – Level 4 PILNA 2015 – Level 6 TERCE 2014 – Level 1 TIMSS 2015 – Intermediate International	Level 2 Level 3 Level 5 Level 2 Intermediate International
Grades 8 & 9	Students demonstrate skills in computation, application problems, matching tables and graphs, and making use of algebraic representations.	PISA 2015 – Level 2 TIMSS 2015 – Low International	Level 2 Intermediate International

To report on math indicator 4.1.1, the following agreement has been reached on what children should be able to do. Tests are expected to be roughly comparable.

Chapter 4. Expectations for 2030 – Setting benchmarks for country-level attainment

Stakeholders have extensively debated how to set minimum proficiency benchmarks; how to help countries set the percentage of students who should be meeting them by 2030; and how to show increases over time. Final and intermediate targets have been difficult to define.

As mentioned earlier, despite individual and cultural variations, speed and accuracy benchmarks could be used as benchmarks for the early grades (but also grade 6 in under developed countries). As identified earlier, 45-60 words per minute reading could be considered a minimum proficiency level for understanding text. Similarly 10-19 correct digit operations per minute could be a minimum calculation proficiency, and writing speed could also be suggested on the basis of research data. (Government officials must understand the rationale behind these and other figures.)

These and other benchmarks are often unattainable by the majority of the population in a country. For example, the US State Department of Education reportedly found that only 59% of students meet basic math performance standards – far short of the 75% goal that was set.[29] One comparison in 2006 of test results in various countries showed low percentages of students getting pass scores (Abadzi, 2006). If illustrated on a normal distribution of ability or advantage, the successful students were outliers scoring 2 or 3 standard deviations above the country mean. This stark gap has been identified in comparisons of PIRLS and TIMSS scores as well.

One hypothesis is that the performance gap between students in poorer and wealthier countries may be partly due to processing speed and its consequences. Students may "know" certain knowledge items but may but not be able to retrieve and process them fast enough to respond. In principle, two students may have the same concepts, classified in approximately similar networks stored within long-term memory. But one student may retrieve items in their memory network and make links to related concepts faster than another. Processing speed is certainly a function of practice, but it also reflects neurological processes. Better off countries may have more students who have been well nourished and stimulated at home, and therefore the 'white matter' in their brains transmit messages faster than that of children who have had less fortunate childhoods.

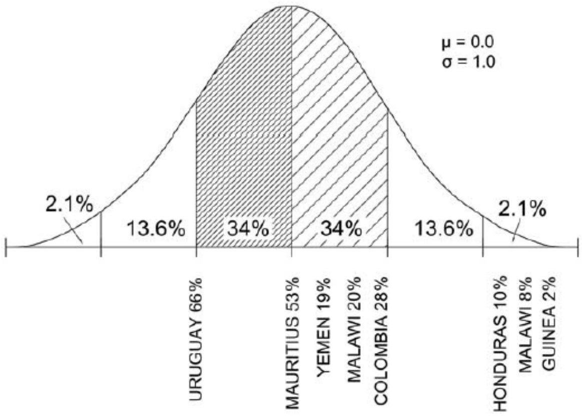
Nevertheless, most students in a country should attain minimum speed parameters. Low-income countries can improve test scores through increasing instructional time, practice, and access to readable textbooks.

Results may have improved in the last decade, but 45-60 words per minute by the end of grade 2, is still attained by only a small percentage of students in many countries (see EGRA results in

[29] This figure is cited in some documents, including a blog http://www.wise-qatar.org/artificial-intelligence-can-save-schools-parents-money-javier-arroyo.

most.) It would be possible to represent the percentages of students meeting benchmarks as areas under a normal curve (in actuality a positively skewed distribution).

Figure 4.1: reading score percentages of satisfactory performance (2006); many countries trained sufficiently only a small minority of students



Sources: Guinea—Barrier et al. 1998; Honduras—Honduras. Secretaría de la Educación 2002; Uruguay – ANEP 2003; Yemen – ERDC 2000; others, Table 1.1. [30]

For example, a country could aim to meet reading benchmarks for 83% of students, i.e. up to 1 standard deviation below average. It would also be possible for a high-income country to attain benchmarks for 96% of those enrolled, i.e. all the way to 2 standard deviations below the mean. (Special education excluded.) A poor country with low outcomes might be realistically able to help only 50% of its students meet benchmarks. An algorithm based on pass percentages of various relevant tests and various socioeconomic indicators could be developed for this purpose. Decimals of standard deviations are a common metric, so it can be interpreted. Thus a reasonably anchored rationale could be established.

And what should be a target improvement percentage from one year to the next, i.e. from 2017 to an intermediate point of 2020 or 2025? Statistical projections can give a rough estimate, but it is also important to inquire what the curricula specify for basic skills instruction and the extent to which they are implemented in various countries. If basic skills are not taught through methods shown to be effective by cognitive neuroscience, improvement may be limited. In fact, there may be no reason to expect improvements.

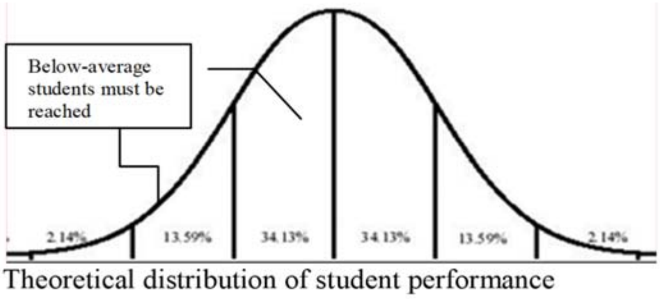
It would thus be possible to specify benchmarks further. Target 4.1 could ensure that every child, regardless of circumstance, completes primary education able to read, write and count well enough to meet minimum learning standards: e.g.

- By 2030, 83% of students worldwide will have attained minimum proficiency in reading and math

[30] Abadzi, 2006. Efficient Learning for the Poor, p. 133.

- By 2030, the proportion of students in country X attaining minimum proficiency will have increased by x% or 0.20 standard deviations (projecting from earlier repeated tests).
- By 2030, the mean proficiency of students will have increased by ZZZ or 10 words per minute over 2017 levels (projecting from earlier repeated tests).

Figure 4.2: Percentages under a normal curve


Below-average students must be reached
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
Theoretical distribution of student performance

Status or progress-based performance? It would be important to include both statements in reporting; i.e., the performance expectations should be stated in terms of status, the proportion of children at the end of primary expected to achieve minimum proficiency in basic skills. There should also be a focus on progress or improvement, i.e., by an intermediate point of 2025 the proportion of students attaining minimum proficiency in will have increased, for example by 10 words per minute or 15 correct single digit calculations, or 5 words written consecutively correctly, etc.

A Worldwide Proficiency Assessment of Numeracy and Literacy?

Treviño & Órdenes (2017) examined the various international tests and developed strategies for their use. One option was to create a Worldwide Proficiency Assessment of Numeracy and Literacy, implemented in the long run. If funding and political will were available, it would be possible to develop and administer a set of new tests, as described in chapters 2 and 3. These could be used to better describe the lower achievement levels in economically deprived populations. A potential process is below.

- A testing schedule every 3-5 years to 2030 for grades 1, 2, 3. To minimize costs, administration in each country would involve the smallest possible representative sample, perhaps 400 students per grade.
- In reading, the tests could include the one-minute letter sounds and connected text of EGRA, as well as subtests from EGMA and/or Number Sets test. Subtests would be chosen for predictive validity of performance in the higher grades, given instruction. Administration time, effort and expenses should be minimized.
- Early-grade tests would be linked to international assessments by giving the tests to students of various grades.

- To monitor progress against a default trend, existing test scores would estimate growth curves to the year 2030. The estimated percentage of students meeting various benchmarks on a default scenario could be compared to the performance aspirations of governments. Countries would be challenged to improve instruction and perform above their projected 'default' performance. Governments would get much specific feedback, along with state-of-the art, science-based advice for improvement. Thoughtful statistical analyses would be necessary to translate results into comprehensible metrics and benchmarks.
- Many of the tests and concepts that could be used have been implemented in academic settings and would require validation in international settings. Thus some applied research is necessary. Researchers currently involved in such studies could potentially receive funding to undertake the needed additional research.
- Staff at UNESCO UIS would propose and monitor tests to be given in various countries, keeping track of contractors who do this, collecting datasets and scheduling and monitoring analyses.

If conducted every 5 years, then 1-minute reading tests and brief math tests could potentially be administered three times until 2030 (2020, 2025, 2030). To measure representative populations of 2400-5800 students in grades 1-3 through oral tests, some guidance may be obtain from EGRA administrations by USAID contractors. Depending on terrain and sample size, these have ranged from $200,000-700,000, but local NGOs and ministries may conduct them for a fraction of the cost. In principle, USAID could be invited to finance them for some countries.

Applied research needed for monitoring target 4.1.1.

Clearly there are gaps in the knowledge of how to monitor and also how to efficiently teach students in grades 1-3 internationally. A preliminary list of research topics is set out below and could be initiated if donors became interested:

- **A words per minute equivalency project**. EGRA datasets are needed for this, particularly of countries where two languages are tested simultaneously. Examples include Kenyan data for English and Swahili, Madagascar data for French and Malagasy, FYROM data for Macedonian and Albanian. A research review is needed, to include involvement with Google or Microsoft for rates of text in pairs of languages. UIS might request the datasets from the Research Triangle Institute or from USAID.
- **Linking PASEC 2nd grade to the EGRA 1-minute connected text words per minute** and EGMA results. PASEC is the only test that is given in grade 2 and 6. LLECE is given in grade 3 and 6; LANA, PIRLS, PILNA in grades 4 and 6. Similar studies could be conducted. (However, PASEC grade 2 may become a paper and pencil test for grade 3 in expectations of linking and harmonizing outcomes of the various international tests.)
- Extension of **Number Sets Test** and its adaptation to international environments.

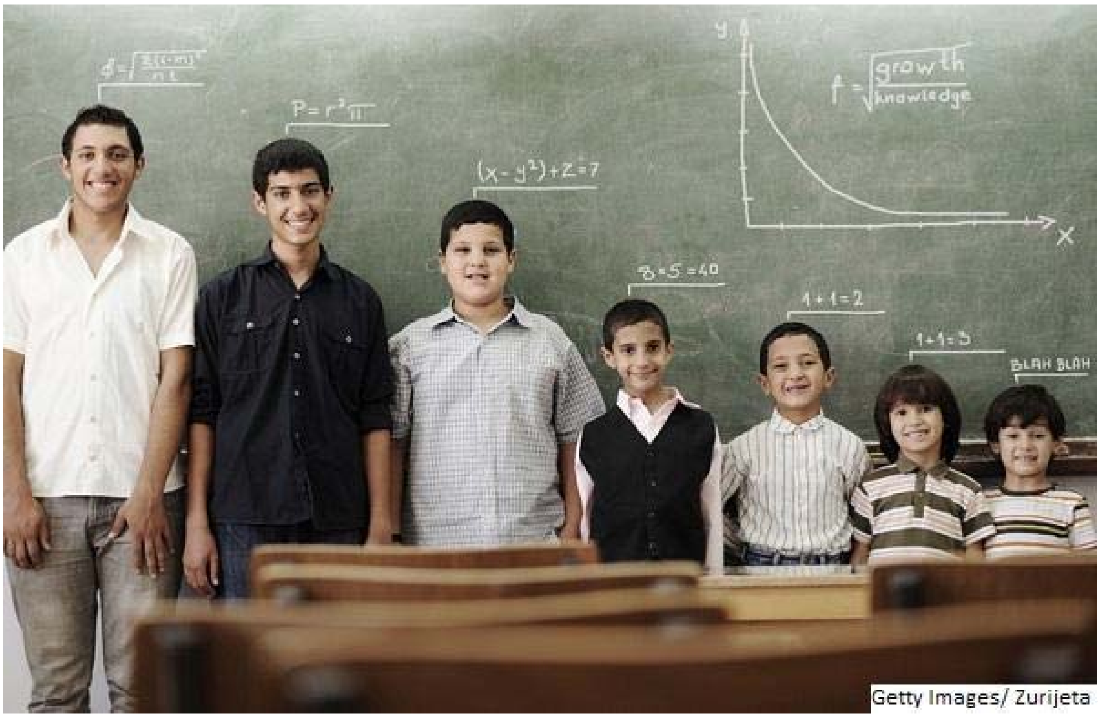
- **Linkage between oral and written language expression and comprehension** that is fluent, instant, and correct. (Language use denotes status matters in addition to actual communication.) There are various tests used worldwide that can be applied, such as the PPVT (Peabody Picture Vocabulary Test). PIRLS and PISA similarly include much conceptual work. Given the similarities of language functions across humans, there could be an international index of correct speech and spelling for native speakers. This would also rate users of formal languages in school.
- **Attention span and other executive functions under academic circumstances**. (These improve with maturation.) For example, the National Foundation of Educational Research in the United Kingdom is developing a baseline test of executive functions and working memory that could be of potential use elsewhere.
- **Linkage between visual complexity and amount of time needed to learn**. The visual complexity scores by graphcom could become a variable in the estimation of nonlinear growth models. Perimetric complexity determines the speed of automaticity, as does the number of items to learn. But the items are rarely loose shapes, they are repetitions of other patterns. It is unclear how these variables combine in memory.

Chapter 5. How will various student groups perform in 2030? Statistical modeling for policy advice

Overall, studies suggest that countries are not making changes quickly enough and many will not meet the goals that have been set by 2030. Can some countries improve on their long-term rankings? To improve outcomes on indicators linked to the SDGs, governments must be aware of likely outcomes in advance and focus on mitigating them. This leads to statistical modeling techniques and estimations of performance growth in a specific timespan. Some essential concepts are below.

Growth models are techniques aimed at estimating a student's capacity for developing a particular skill or ability. They are used to estimate the highest likely performance after tutoring, for example. Many studies have been conducted to facilitate dyslexia and dyscalculia in high-income countries. The early-grade research relies on those.

A growth to standards illustration is shown below (Figure 5.1)

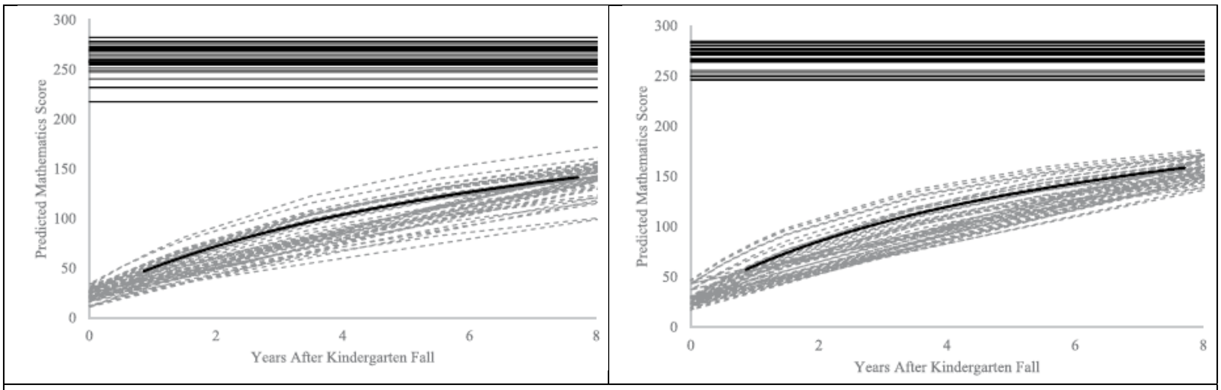


Getty Images/ Zurijeta [31]

One example, derived from special education is dynamic measurement modeling, which can measure "growth to standards" through the modeling of longitudinal testing data without the need for extensive one-on-one testing (McNeish & Dumas, 2017; Dumas & McNeish, 2017).

[31] https://edexcellence.net/articles/why-states-should-use-student-growth-and-not-proficiency-rates-when-gauging-school.

Figure 5.2: An example of a dynamic nonlinear growth curve (Dumas & McNeish, 2017)



Compared to performance at full capacity, it is possible to estimate the ability of various groups to perform a number of years after training, as well as estimate the unrealized available potential (curve to the right)

Growth modeling has become sophisticated in the United States, where teachers are evaluated for accountability reasons. The paradigm typically focuses on assessing the ‘value added’ of teacher instruction given students’ prior knowledge, background characteristics, characteristics of their peers and teachers’ previous records. A single pretest and post-test is used. Such models could (a) predict student performance in various countries for 2030 and (b) help develop benchmarks at various ages for various countries. Linear and nonlinear modeling would be relevant (e.g. use of quantile regression for nonlinear modeling).

The paradigm can be expanded to the monitoring purposes of the target 4.1. Particularly relevant is a “growth-to-standards” methodology, which projects likely scores to a point in time. "Growth to standards" models could project for each country performance at various levels to 2030, given a default current trend. For example, the 2030 EGRA scores may be projected for average and lower-quartile students on the basis of 2006 and 2013 scores. Likely performance estimates could be projected for various countries to 2030 under a “default” scenario, as well as under scenarios of improved learning conditions. Gaps are likely to be shown. Then Ministries of Education could try to improve instruction to exceed the projections and "move the needle" to close the gap until 2030. The estimation of future attainment may add value by notifying countries where they will likely be in 12 years.

One implication of growth modeling is that a country can be compared to itself and its targets to 2030. Thus, international tests and tests pertinent to that country could be used. Standards should be relatively homogeneous across countries. (Note that this is easier in the earlier, rather than in the later, grades). The main indicator would be percentage of students falling below certain benchmarks.

The growth standards in grades 1-3 would be based on automaticity indicators (such as words per minute or math operations per minute). Given their role in complex cognition, such indicators should be 'non-negotiable' for all countries. Nearly all students would be expected to score above the benchmark. The improvement rate in each country could be determined on the basis of projections to 2030 and efficient instructional methods.

From grade 4 onwards, the international and regional tests or various national test benchmarks could be most useful. Modeling would determine the percentage students likely to meet various benchmarks by 2030 by default. Then countries could estimate/predict the progress rate and the percentage of students meeting these benchmarks on an improved scenario (and better instruction).

In addition trajectories may be used to compare learning from a broader perspective across countries. For example, a given country may be effective in generating academic growth in the population, but that population may enter school at a comparatively low level of achievement, putting that country at risk for being viewed negatively in international comparisons. By contrast, another country may have a population that enters school with relatively high achievement already on average, and therefore does not actually generate as much learning as another country, despite being compared favorably. In this way, single-administration achievement test scores privilege countries with populations who enter schools with relatively high achievement already. But the model allows for the fair comparison of learning across countries, without needing to change the data being collected.

Data to develop predictions already exist for many countries. The number of assessments and test specifications vary for each country. Better off countries have international assessments that make it possible to assess percentages of students who perform at a minimal, sufficient, and advanced level (e.g. Finland). Then governments can set benchmarks to be attained for various population segments. Poorer countries often lack data from PIRLS, TIMSS, or PISA , but they may have outcomes from PASEC, SACMEQ, or LLECE. They may also have data from repeated EGRA and EGMA tests, as for example in Gambia. Multiple test administrations help to create a "Rosetta stone" scenario, where test scores can be imputed in tests a country has not taken.

The ground for such work has been set. In 2017, UNESCO financed an exploration of data-linking methodologies. This database is part of World Bank and UNESCO Institute for Statistics (UIS) partnership to advance this effort. Altinok et al. (2018) developed a model to harmonize available learning data across different types of international and regional assessments.

The goal of the Altinok et al. (2018) study was to provide a practical, yet rigorous and globally comparable, set of estimates with large and inclusive country coverage over time. The method is preliminary and has imperfections that result in some unexpected rankings, e.g. placing Kenya above Chile in some measures. However, as more countries join international and regional assessments, and do so for longer, the accuracy and robustness of the harmonization exercise will improve. But, although mean scores may vary by linking methods, ranks and relative performance are rather robust. This gives the default trend given 'business as usual' and should generate reasonable predictions to 2030. It could also answer the question of how many years it may take for various countries to rise to desired performance levels at current growth rates.

For multiple tests of the same type, the researchers computed performance using the growth rate, for example, between PIRLS 2001 and 2006 (estimating a linear or nonlinear solution). Altinok et al., used the three different benchmarks provided by TIMSS, PIRLS, and PISA: minimum, intermediate, advanced (the benchmarks of 400, 475, and 625 respectively in reading, math, and science).

A first analysis of this dataset revealed a few important trends: (a) Learning outcomes in developing countries often cluster at the bottom of a global scale; (b) Although variation in performance is high in developing countries, the top performers still often perform worse than the bottom performers in developed countries; (c) Gender gaps are relatively small, with high variation in the direction of the gap by region. At the primary level, mean scores fluctuate over time, but overall they have increased. Annual growth rates vary from 0.10 to 0.62 percent. (The largest gains are in Hong Kong, Iran, and Finland.) At the secondary level, performance has decreased overall, due to low-performing students remaining longer in the system. Distributions reveal meaningfully different trends than mean scores, with less than 50 percent of students reaching the minimum global threshold of proficiency in developing countries relative to 86 percent in developed countries.

There have been a few other attempts to harmonize datasets, notably the Swedish LYNC project (Strietholt & Rosén, 2016). The LINCS project aims to calibrate the achievement scales from past and present studies on a common Item-Response-Theory (IRT) scale. This delivers an empirical basis for investigating the long-term effects of educational policy and policy-related issues on educational outcomes. When looking at reading at the end of primary school, for instance, trend studies with PIRLS (Progress in International Reading Literacy Study) data are limited in the covered time span, because the study was first implemented in 2001. Limited analyses and publications were carried out by late 2018.

Issues in estimating performance gaps to 2030

The challenges of creating a common scale, and common achievement levels through an equating process for 2030, are enormous. There are differences in the content and skills assessed in each of the studies. Trends are not always monotonic, error sizes and biases would increase. Nevertheless, assumptions about the commonalities of human skills may help. Nevertheless, a "Rosetta stone" exercise has already been conducted. Countries taking certain regional tests have received putative scores in international tests. Oral tests, and potentially others discussed in this document, can be included.

It is important to find means that enable countries to take action. As discussed earlier, mere general statistics offer insufficient advice. Specific findings must be applied, that come from neurocognitive research. Some academics have thought a lot about early-grade variables predictive of later performance and have developed complex models (e.g., Bull et. al., 2009; Chu et al., 2016; Koponen et al., 2007, Peng et al., 2016, Reigosa et al., 2013).

Other possibilities exist for projecting likely performance to 2030. Human cognition makes it possible to establish a few global predictive variables (Turchin et al., 2017). The advent of machine-learning algorithms and "big data", i.e. hundreds of variables related to students, has created much predictive capacity for future test scores (e.g. Iqbal et al., 2017). There is considerable educational research focused on predicting student performance on the basis of information-processing variables. In principle performance at country level at the end of primary school could be predicted by a model that includes variables such as perimetric complexity of letters, orthographic consistency, diglossia, instructional time, practice amounts, teacher effectiveness measures, textbook availability, and various income- and teacher-related variables.

It would be important to test the extent to which such models can reliably predict student performance in various countries for 2030 and also offer concrete suggestions to modify early-grade curricula in ways that little-educated teachers in poorly resourced classrooms can implement. The World Bank could possibly finance a pilot study to test the feasibility of developing "growth to standards" linear and non-linear curves.

Policy dialogue on performance by 2030: The case of the Seychelles

Seychelles is an island country of about 91,000 inhabitants in the Indian Ocean. It is classified as high-income, with much revenue from tourism and fishing.

The government is very interested in improving its learning outcomes further and becoming internationally competitive. It therefore invited UNESCO-IBE, an institute particularly interested in cognitive science, to bring in relevant expertise. A team of measurement and learning specialists, therefore, carried out a series of missions.

The citizens of the Seychelles speak a French Creole that is spelled transparently. Curricula specify instruction in Creole in the first three grades, with English used thereafter. But much early instruction, including for reading, takes place in English. The country participates in SACMEQ and administers its own national tests in English. In the 2013 administration of SACMEQ IV (the results of which had not been fully analyzed by 2018), Seychelles had the highest score in reading and the third highest score in math (Table 5.1). In principle, these results are very encouraging for the country's progress in basic skills acquisition.

Table 5.1: SACMEQ IV 2013 average results, unpublished

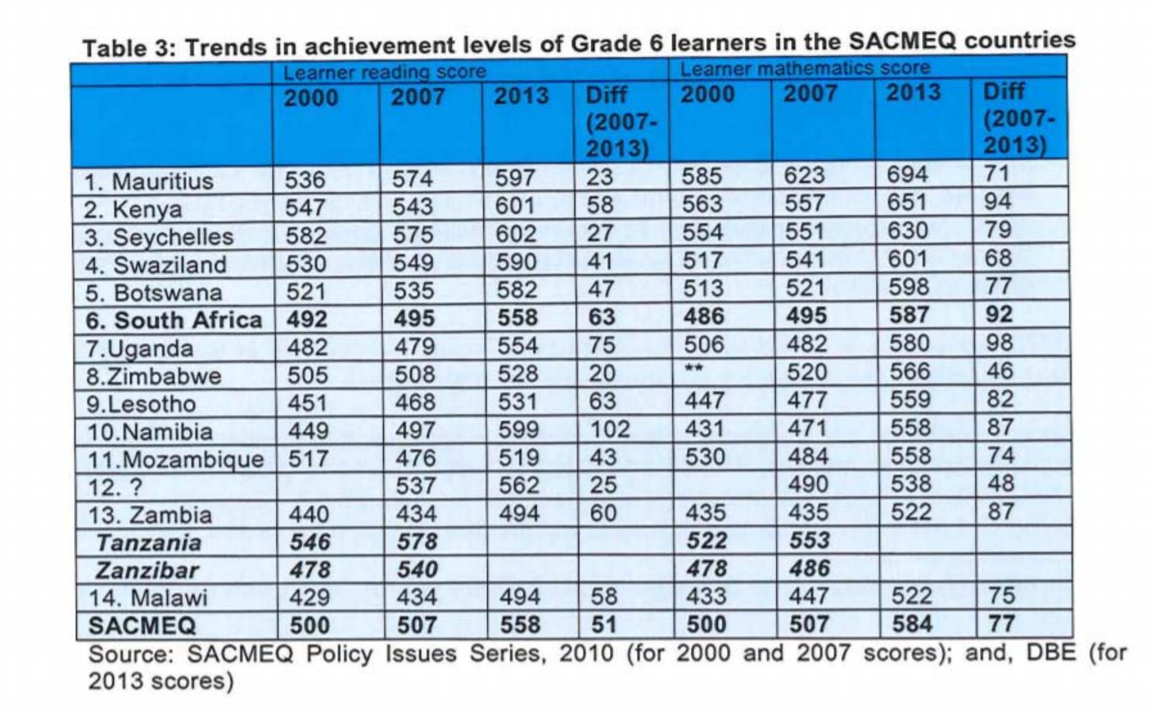


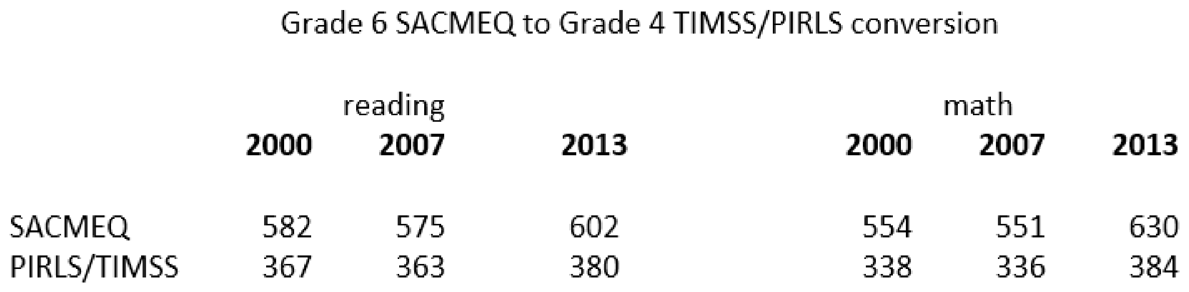
Table 3: Trends in achievement levels of Grade 6 learners in the SACMEQ countries

	Learner reading score				Learner mathematics score			
	2000	**2007**	**2013**	**Diff (2007-2013)**	**2000**	**2007**	**2013**	**Diff (2007-2013)**
1. Mauritius	536	574	597	23	585	623	694	71
2. Kenya	547	543	601	58	563	557	651	94
3. Seychelles	582	575	602	27	554	551	630	79
4. Swaziland	530	549	590	41	517	541	601	68
5. Botswana	521	535	582	47	513	521	598	77
6. South Africa	**492**	**495**	**558**	**63**	**486**	**495**	**587**	**92**
7.Uganda	482	479	554	75	506	482	580	98
8.Zimbabwe	505	508	528	20	**	520	566	46
9.Lesotho	451	468	531	63	447	477	559	82
10.Namibia	449	497	599	102	431	471	558	87
11.Mozambique	517	476	519	43	530	484	558	74
12. ?		537	562	25		490	538	48
13. Zambia	440	434	494	60	435	435	522	87
Tanzania	***546***	***578***			***522***	***553***		
Zanzibar	***478***	***540***			***478***	***486***		
14. Malawi	429	434	494	58	433	447	522	75
SACMEQ	**500**	**507**	**558**	**51**	**500**	**507**	**584**	**77**

Source: SACMEQ Policy Issues Series, 2010 (for 2000 and 2007 scores); and, DBE (for 2013 scores)

The Altinok et al. study had information linking SACMEQ data to TIMSS and PIRLS. Linkage gave a more realistic picture of possible scores if students of the Seychelles took these tests.

Table 5.2: Conversion of SACMEQ scores to TIMSS and PIRLS 2011[32]



Grade 6 SACMEQ to Grade 4 TIMSS/PIRLS conversion

	reading			math		
	2000	**2007**	**2013**	**2000**	**2007**	**2013**
SACMEQ	582	575	602	554	551	630
PIRLS/TIMSS	367	363	380	338	336	384

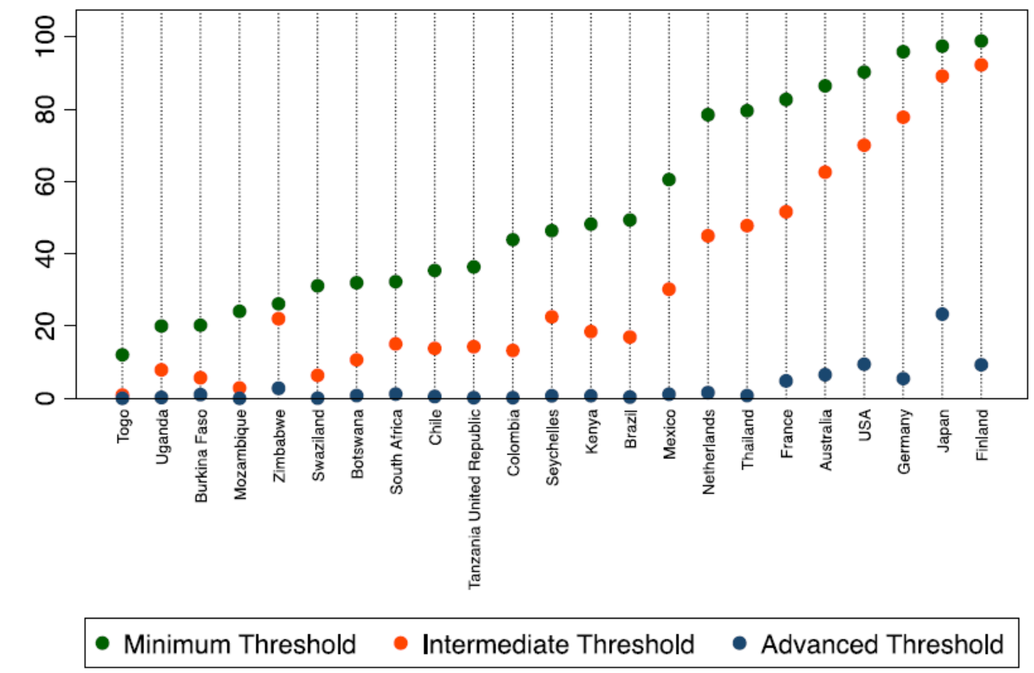
Translated into international assessments, the Seychelles 6th graders would perform near the bottom of the PIRLS scale. Linked scores, furthermore, show that about 35% perform below a minimum performance level; about 40% perform at the minimum level and another 20% perform

[32] The conversion rates used were: for mathematics between SACMEQ and TIMSS: 0.630929; for reading between SACMEQ and PIRLS: 0.6093 (N. Altinok, personal communication).

at the intermediate level. Only about 2% perform at the advanced level (Figure 5.1). These results suggest much potential for improvement.

Figure 5.1: African student performance levels in linked tests (Altinok et al., 2018)

Figure 6b: Percent of Students Achieving Low, Intermediate and Advanced Average Primary



The Seychelles could attain an average score of 500 in TIMSS and PIRLS by 2030 and could aim for the vast majority of its students to perform at least at the minimum level by then. An example towards which to aspire in the intermediate years is Malta. This island nation is in some respects comparable. Malta uses the Maltese language (Maghrebine Arabic written in the Roman script) along with English. In 2011, it attained scores of 477 and 496 in PIRLS and TIMSS respectively. Seychelles could prepare for this goal in target 4.1.

At the heart of the Seychelles performance are the country's linguistic complexities. To score highly in PIRLS, 4th graders must read passages of 800-1000 words and answer a set of short-answer and multiple-choice questions in 20 minutes. This could be accomplished in Creole, just as Malta teaches in Maltese. However, prima facie, few students can carry out this task in Creole or in English. Fluent and fast reading is an important prerequisite for many skills. Instruction must focus on early-grade processing speed that will enable 4th graders to undertake the above task.

To gauge student performance in this task, the Ministry of Education may use one of the released texts and administer it to 4th graders, monitoring the time they take to read it and the precision of comprehension. Depending on outcomes, the MOE may (a) include more reading practice and language preparation early in the curricula and may (b) remediate students who fall below minimum performance.

After improvements, the government could choose to administer PIRLS in Creole and English. Until then, and for closer monitoring, national tests could be linked to international tests through the addition of a linking module. This would enable item statistics of Seselwa students to be compared to those of other countries. The important issue, however, is improvement.

In principle, it would be possible to use SACMEQ scores and estimate likely growth curves to 2030, after SACMEQ analyses are complete. Also national test scores could be used for the process. In the meantime, the harmonized database that was sponsored by UIS and the World Bank has provided important insights to help the government make decisions regarding its curricula and citizen skills expected for 2030. Thus policy dialogue can take place in other countries as well.

References

Abadzi, H. 2015. Training the 21st Century Worker: Policy Advice from the Dark Network of Implicit Memory. UNESCO: IBE Working Papers on Curriculum Issues Nº 16. http://www.ibe.unesco.org/en/document/training-21st-century-worker-ibe-working-papers-curriculum-issues-n%C2%B0-16.

Abadzi, H. (2017). Turning a molehill into a mountain? Why reading programs are failing the poor worldwide. Prospects, 177. DOI: 10.1007/s11125-017-9394-9; http://rdcu.be/qqMe.

Abadzi, H. 2013. Developing Cross-Language Metrics for Reading Fluency Measurement: Some Issues and Options. World Bank: Global Partnership for Education Working Paper Series No. 6.

Altinok, N., Angrist, N., & Patrinos, H. A. (2018). Global Data Set on Education Quality (1965–2015). World Bank: Policy Research Working Paper 8314.

Anderson, J. R., Pyke, A. a., & Fincham, J. M. 2016. Hidden Stages of Cognition Revealed in Patterns of Brain Activation. Psychological Science, 2016; DOI: 10.1177/0956797616654912.

ASER, 2017. Beyond the Basics: A survey for rural Indian youth. 4-page summary, 2018.

Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014). State and trait effects on individual differences in children's mathematical development. Psychological science, 0956797614547539.

Baddeley, A., M.W. Eysenck and M.C. Anderson. 2015. Memory. Second edition. New York: Psychology Press.

Bull, R., Espy, K., & Wiebe, S. A. 2009. Short-Term Memory, Working Memory, and Executive Functioning in Preschoolers: Longitudinal Predictors of Mathematical Achievement at Age 7 Years. Dev Neuropsychol. 2008; 33(3): 205–228. DOI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2729141/.

Buss, D. 2015. Evolutionary Psychology 5th Edition, Routledge.

Chang, Li-Yun, Chen, Yen-Chi & Perfetti, C. 2017. GraphCom: A multidimensional measure of graphic complexity applied to 131 written languages. Behavioral Research Methods, 2017. DOI: 10.3758/s13428-017-0881-y.

Cheng, Y.-L. & Mix, K. S. 2014. Spatial training improves children's mathematics ability. Journal of Cognition and Development, 15, 1.

Chu, F. W., vanMarle, K., & Geary, D. C. 2016. Predicting Children's Reading and Mathematics Achievement from Early Quantitative Knowledge and Domain-General Cognitive Abilities. Front Psychol. 2016 May 25;7:775. DOI: 10.3389/fpsyg.2016.00775. eCollection 2016.

Cooper, G. and J. Sweller. 1987. "The effects of schema acquisition and rule automation on mathematical problem-solving transfer." Journal of Educational Psychology 79, 347–362.

Crossland, M. D., Legge, G. E. & Dakin, S. C. 2008. The development of an automated sentence generator for the assessment of reading speed. Behavioral and Brain Functions 4:14. doi:10.1186/1744-9081-4-14. [DGP]

Coltheart, M. & Crain, S. Are there universals of reading? We don't believe so. Behavior and Brain Sciences (2012).

Cui, J. Zhang, Y., Cheng, D., Li, D., & Zhou, X. 2017. Visual Form Perception Can Be a Cognitive Correlate of Lower Level Math Categories for Teenagers. Frontiers in Psychology, doi: 10.3389/fpsyg.2017.01336.

Dehaene, S. 1997. The Number Sense. Oxford.

Dehaene, S., & Cohen, L. (011. The unique role of the visual word form area in reading. Trends in Cognitive Sciences, 15(6), 254-262.

DeSouza, D. 2010 (Ed). Mind Brain and Education. Bloomington, IN: Solution Tree Press, p. 187.

Dumas, Denis G. Dumas & Daniel M. McNeish. 2017. Dynamic Measurement Modeling: Using Nonlinear Growth Models to Estimate Student Learning Capacity. Educational Researcher, Vol. 46 No. 6, pp. 284–292. DOI: 10.3102/0013189X17725747.

Evans. T. M., Flowers, D. L., Luetje, M. M., Napoliello, E., Eden, G. F. 2016. Functional neuroanatomy of arithmetic and word reading and its relationship to age. NeuroImage, 143, 304–315. http://dx.doi.org/10.1016/j.neuroimage.2016.08.048.

Fedzedchkina, M., Chu, B., Jaeger, T.F. 2017. Human Information Processing Shapes Language Change. Psychological Science 1–11. DOI: 10.1177/0956797617728726.

Feigenson, L., Libertus, M. & Halberda, J. 2013. Links Between the Intuitive Sense of Number and Formal Mathematics Ability. Child Development Perspectives, DOI: 10.1111/cdep.12019.

Fernandes, T., Araújo, S., Sucena, A., et al. 2017. The 1-min Screening Test for Reading Problems in College Students: Psychometric Properties of the 1-min TIL. Dyslexia. DOI: 10.1002/dys.1548.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Karns, K. 1998. High-achieving students' interactions and performance on complex mathematical tasks as a function of homogeneous and heterogeneous pairings. American Educational Research Journal, 35, 227-268.

Geary, D. C., Hoard M. K., Nugent L., & Bailey D. H. 2013. Adolescents' Functional Numeracy Is Predicted by Their School Entry Number System Knowledge. PLoS ONE. 8(1).

Geary, D. C., Bailey, D. H., Hoard, M. K. 2009. Predicting Mathematical Achievement and Mathematical Learning Disability With a Simple Screening Tool: The Number Sets Test. J Psychoeduc Assess. 2009 Jun; 27(3): 265–279. DOI: 10.1177/0734282908330592.

Geary, D. C., vanMarle, K., Chu, F., Rouder, J, Hoard, M. K., & Nugent. 2018. Early conceptual understanding of cardinality predicts superior school-entry number system knowledge. Psychological Science, 29.

Geary DC, Hoard MK, Byrd-Craven J, Nugent L, Numtee C. 2007. Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. Child Development. 2007;78:1343–1359.

Geary, D., C., vanMarle, K., Chu, F. W., Rouder, J., Hoard, M. K., & Nugent, L. 2017. Early Conceptual Understanding of Cardinality Predicts Superior School-Entry Number-System Knowledge. Psychological Science, 1–15, DOI: 10.1177/0956797617729817.

Geary, D.C., Hoard, M. K., Nugent, L., & Bailey, H. D. 2013. Adolescents' functional numeracy is predicted by their school entry number system knowledge. PLoS ONE, 8(1): e54651.

Geary, D. C. 2011. Cognitive predictors of individual differences in achievement growth in mathematics: A five-year longitudinal study. Developmental Psychology, 47, 1539-1552.

Grabner, R.H., Saalbach, H. & Eckstein, D. 2012. Language switching costs in bilingual mathematics learning. Mind, Brain and Education, 6(3), 147-155.

Hartley, Michael J.; Swanson, Eric V.; 1986. Retention of basic skills among dropouts from Egyptian primary schools. World Bank, Working Paper (Numbered Series), Report Number EDT40.

Hasbrouck, J. & Tindal, R. 2006. Oral reading fluency norms: A valuable tool for assessment. Reading Teacher, 59 (7), 636-644.

Hinault, T. & Lemaire, P. 2016. What does EEG tell us about arithmetic strategies? A review. International Journal of Psychophysiology.

Iqbal, Z., Qadir, J., Mian, A. N. & Kamiran, F. 2017. Machine Learning Based Student Grade Prediction: A Case Study. https://arxiv.org/pdf/1708.08744.pdf.

Kirby, J., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? Literacy, 42, 75-82.

Koponen, T., Aunola, K., Ahonen, T., & Nurmi, J. E. 2007. Cognitive predictors of single-digit and procedural calculation skills and their covariation with reading skill. J Exp Child Psychol. 2007 Jul;97(3):220-41.

Lauer, J. E. & Lourenco, S. F. 2016. Spatial Processing in Infancy Predicts Both Spatial and Mathematical Aptitude in Childhood. Psychological Science, 2016; DOI: 10.1177/0956797616655977.

Strietholt ck, N. 2018. Kichwa or Quichua? Competing Alphabets, Political Histories, and Complicated Reading in Indigenous Languages. Comparative Education Review.

Looi, C. et al. 2016, The Neuroscience of Mathematical Cognition and Learning, OECD Education Working Papers, No. 136, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jlwmn3ntbr7-en.

Leibovich, T., Katsin, N., Harel, M., & Henik, A. 2016. From ‘sense of number’ to ‘sense of magnitude’ – The role of continuous magnitudes in numerical cognition. Behavioral and Brain Sciences, 2016; 1 DOI: 10.1017/S0140525X16000960.

Masson, M. E. J. 1983. Conceptual processing of text during skimming and rapid sequential reading. Memory & Cognition 11(3):262–74. [DGP]

Muroya, N., Inoue, T., Hosokawa, M., Georgiou, G., Maekawa, H., & Parrila, R. 2017. The role of morphological awareness in word reading skills in Japanese: A within-Language cross-orthographic perspective. Scientific Studies of Reading, 21(6).

McKoon, G. & Ratcliff, R. 2017. Adults with poor reading skills and the inferences they make during reading. Scientific Studies of Reading, 21,4, 292-309.

Martirossian, J. & Lewin, D. 2001. Decision Making in Communities: Why Groups of Smart People Sometimes Make Bad Decisions. Community Associations Institute.

Mou, Y., Li Y., Hoard M. K., Nugent L. D., Chu F. W., Rouder J. N., & Geary D. C. 2016. Cognitive Development, 39, 141-153.

Nath, S., & Szücs, D. 2014. Construction play and cognitive skills associated with the development of mathematical abilities in 7-year-old children. *Learning and Instruction, 32,* 73-80. http://dx.doi.org/10.1016/j.learninstruc.2014.01.006.

Parker, R. I., Tindal, G., & Hasbrouck, J. 1991. Progress monitoring with objective measures of writing performance for students with mild disabilities. Exceptional Children, 58, 61–73.

Percival G. Matthews, Mark Rose Lewis, Edward M. Hubbard. 2015. Individual Differences in Nonsymbolic Ratio Processing Predict Symbolic Math Performance. Psychological Science.

Menon V, Qin S, Cho S, Chen T, Rosenberg-Lee M, Geary. 2014. Hippocampal-neocortical functional reorganization underlies children's cognitive development. Nature Neuroscience. 2014.

Mizuno, K., Tanaka, M.,Yamaguti, K., Kajimoto, O., et al. 2011. Mental fatigue caused by prolonged cognitive load associated with sympathetic hyperactivity. Behavioral and Brain Functions, 7,17.

Pelli, D. G., Chung, S. T. L., Legge, G. E. 2012. Theories of reading should predict reading speed. Journal: Behavioral and Brain Sciences, 35(5) doi.org/10.1017/S0140525X12000325.

Perfetti, C., Cao, F., & Booth, J. 2013. Specialization and universals in the development of reading skill: How Chinese research informs a universal science of reading. Scientific Studies of Reading, 17(1), 5–21.

Peng, P., Namkung, J. M., Fuchs, D., Fuchs, L. S., Patton, S., Yen, L., et al. 2016. A longitudinal study on predictors of early calculation development among young children at risk for learning difficulties. J Exp Child Psychol. 2016 Dec;152:221-241. DOI: 10.1016/j.jecp.2016.07.017.

Price, G. R., Mazzocco, M. M. M. & Ansari, D. 2013. Why Mental Arithmetic Counts: Brain Activation during Single Digit Arithmetic Predicts High School Math Scores. Journal of Neuroscience, 33 (1): 156. DOI: 10.1523/JNEUROSCI.2936-12.2013.

Qin, S., Cho, S., Chen, T., Rosenberg-Lee, M., Geary, D. C., & Menon, V. 2014. Hippocampal-neocortical functional reorganization underlies children’s cognitive development. Nature Neuroscience, 17, 1263-1269.

Reigosa-Crespo, V., Torres, R., Mosquera, R., et al (in preparation). Subitizing predicts individual variability of math competence in older learners.

Reigosa-Crespo V, González-Alemañy E, León T, Torres R, Mosquera R, et al. 2013. Numerical Capacities as Domain-Specific Predictors beyond Early Mathematics Learning: A Longitudinal Study. PLoS ONE 8(11): e79711. DOI:10.1371/journal.pone.0079711.

Research Triangle Institute. 2012. Performance of students in reading and mathematics, pedagogical practices and school management in Dukkala Abda, Morocco. USAID.

Research Triangle Institute. 2010. Ethiopia Early Grade Reading Assessment. Data Analytic Report: Language and Early Learning. Ethiopia Early Grade Reading Assessment. Ed Data II Task Number 7 and Ed Data II Task Number 9. October 31, 2010.

Rovee-Collier, C., Hayne, H., & Colombo, M. 2000. The Development of Implicit and Explicit Memory. Advances in Consciousness Research.

Rueckl, J. G., Paz-Alonso, P. M., Molfese, P. J., et al. 2015. Universal brain signature of proficient reading: Evidence from four contrasting languages. PNAS 2015 112 (50) 15510-15515.

Seymour, P., H.K.M. Aro, and J.M. Erskine. 2003. Foundation Literacy Acquisition in European Orthographies. British Journal of Psychology 94, no. 2: 143–174.

Siegler, R. S., Duncan, G. J.., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel et al. 2012. Early Predictors of High School Mathematics Achievement. Psychological Science; DOI: 10.1177/0956797612440101.

Siegler, R.S. 2009. Improving the numerical understanding of children from low-income families. Child Development Perspectives, 3(2), 118-124.

Skagenholt, M,, Träff, U., Västfjäll, D., & Skagerlund, K. (2018) Examining the Triple Code Model in numerical cognition: An fMRI study. PLoS ONE 13(6): e0199247. https://doi.org/10.1371/journal.pone.0199247.

Strietholt, Rolf & Monica Rosén. 2016. Measurement: Interdisciplinary Research and Perspectives, 14, 1-26.

Supekar, K., Swigart, A. G., Tenison, C., Jolles, D.D. Rosenberg-Lee, M., Fuchs, L., & Menon, V. 2013. Neural predictors of individual differences in response to math tutoring in primary-grade school children. www.pnas.org/cgi/doi/10.1073/pnas.1222154110.

Treviño. E. & Órdenes, M. 2017. Exploring Commonalities and Differences in Regional and International Assessments. UNESCO Institute of Statistics, Information Paper No. 48.

Turchin, P. Currie, T. E., Whitehouse, H. el al. 2017. Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. PNAS. www.pnas.org/cgi/doi/10.1073/pnas.1708800115.

Yu, C. Watanabe, T., Sagi, D., & Levi, D. 2009. Perceptual learning: Functions, mechanisms, and applications. Vision Research 49, 2531–2534.

UNESCO. 2015. Education for All 2000-2015: Achievements and Challenges. Paris, UNESCO Publishing.

Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K. et al. 2013. Deconstructing Building Blocks: Preschoolers' Spatial Assembly Performance Relates to Early Mathematics Skills. Child Development, 85(3), 1062-1076.

Vista, A., Kim, H., & Care, E. 2018. Use of data from 21st century skills assessments: Issues and key principles. Brookings Institute.

Wang, J. Odic, D., Halberda, J., & Feigenson, L. 2016. Changing the precision of preschoolers' approximate number system representations changes their symbolic math performance. Journal of Experimental Child Psychology, 2016; 147: 82. DOI: 10.1016/j.jecp.2016.03.002.

Watts, T. W., Duncan, G. J., Clements, D. H., & Sarama, J. 2017. What Is the Long-Run Impact of Learning Mathematics During Preschool? Child Development.

Weber, A. M., Fernand, L., C., Galasso, E. et al. 2015. Performance of a receptive language test among young children in Madagascar. PLoS One 10 (4), e0121767. 2015 Apr 01.

Whiteley, W., Sinclair, N. and Davis, B. 2015. What is spatial reasoning? In Davis, B. and the Spatial Reasoning Group (Eds). Spatial Reasoning in the Early Years. Principles, Assertions, and Speculations. Abingdon: Routledge.

Wolfgang C. H., Stannard, L. L.,& Jones I. 2001. Block play performance among preschoolers as a predictor of later school achievement in mathematics. *J. Res. Child. Educ.* 15 173–180 10.1080/02568540109594958.

Zhang, J., Fan, X, Kwing, S. C., Meng, Y., Cai, Z., & Bi, Y. H. 2017. The role of early language abilities on math skills among Chinese children. PLoS One. 2017; 12(7): e0181074. 10.1371/journal.pone.0181074.

Zhu, M., Cai, D., Leung, A. W. S. 2017. Number Line Estimation Predicts Mathematical Skills: Difference in Grades 2 and 4. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2017.01576.