All Children Reading – Asia (ACR – Asia)

EGRA Benchmarks and Standards Research Report

Submission Date: December 21, 2017

AID-OAA-TO-16-00017 Number: REQ-ASIA-16-00017
Activity Start Date and End Date: September 30, 2016, to September 29, 2021
TOCOR: Mitch Kirby

Submitted by: **RTI International**
3040 Cornwallis Road
Research Triangle Park, NC 27709-0155
Tel: (919) 541-6000

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AIR	American Institutes for Research
ARMM	Autonomous Region of Muslim Mindanao
DepEd	Philippine Department of Education
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
EGRA	Early Grade Reading Assessment
LTTP	Liberia Teacher Training Program
NCD	Papua New Guinea National Capital District
ORF	oral reading fluency
PEARL	Pacific Early Age Readiness and Learning
PRP	Pakistan Reading Project
QRP	Quality Reading Project
READ TA	Reading for Ethiopia's Achievement Developed Technical Assistance
SDG	Sustainable Development Goal
UNESCO	United Nations Educational, Scientific and Cultural Organization
USAID	United States Agency for International Development
WHP	Papua New Guinea Western Highlands Province

1 Executive Summary

The Early Grade Reading Assessment (EGRA) is widely used to assess reading proficiency in developing countries. Benchmarks were introduced to simplify EGRA results into a single indicator against which countries could measure their children's reading progress. This report is about the process, rational and considerations for setting benchmarks (a desired level of performance on a reading task) and targets (the percent of students intended to reach the performance level). The report addresses three purposes of benchmarking:

1. Country purpose: Track progress in reading within a national education system and provide data to inform efforts to improve education quality.
2. Agency purpose: Help USAID monitor the progress of projects and countries working to reach specific children-reading goals.
3. The global purpose: Measure progress in reading for all; provide a basis for global advocacy; and provide a method of assessing Indicator 4.1.1 of the Sustainable Development Goals

1.1 Development and Measurement of Reading

Development of Reading

Reading development is conceptualized as an integrated series of skills: Accuracy is an important predictor of comprehension in the early stages of reading but then reading speed and fluency takes over as a better predictor once a child surpasses basic reading ability

Reading fluency across languages

The rate of fluency acquisition varies across languages. In theory, the strength of the relationship between fluency and comprehension could vary with orthographic depth because it is possible to read shallow orthographies without comprehension. However, a moderate-strong relationship between fluency and comprehension has been found in languages with a range of orthographies. In general, evidence supports a universal theory of learning to read applicable across all languages.

Rationale for measuring fluency

For students in early grades the goal is to read fluently with comprehension. Comprehension is difficult to assess reliably but fluency acts as a proxy for comprehension and thus is the focus of benchmarking. In addition, fluency is an important skill in its own right, it is straightforward to measure and is a transparent measure readily understood by parents and teachers.

Rationale for creating benchmarks

There are several reasons to set benchmarks. First, benchmarks allow education systems to articulate their definition of reading proficiency. Second, they communicate this definition of reading proficiency and provide a standard for others to aim for. Third, benchmarks result in a count of the number of students with reading proficiency. Counts are more readily interpreted, especially by non-experts. Counts can also be summed across contexts, for example to assess progress against the USAID All-Children-Reading targets.

Recommendations

1. EGRA assessments and fluency benchmarks are appropriate to use in Asian languages, based on the science of reading
2. Language-specific benchmarks should be set, because orthographies and other factors influence the rate of fluency acquisition.

3. Fluency assessments need considered approaches to counting words in languages with ambiguous word boundaries including: counting characters or syllables rather than words, focusing on errors in word segmentation rather than accurate word segmentation, or convening experts to adjudicate on the count of words.
4. Tentative conclusions from research to date are that oral reading fluency is a good proxy for comprehension across languages but this assumption should be tested when working in new languages
5. Fluency should be assessed separately in different language forms – such as with and without diacritics

1.2 Fluency Benchmarking Data

1.3 Using Data to Set Fluency Benchmarks

To understand the role of data in setting fluency benchmarks, it helps to restate the goals of fluency benchmarks. They include:

- Defining proficiency in reading fluency in a given language and education system
- Providing a goal for students and educators
- Providing a means to track progress of an education system

It is possible that a fluency benchmark set without using data can fulfill these functions. A benchmark set, for example, using the judgement of a panel of experts may offer a credible definition of proficiency and a useful means to motivate and track educational progress. In our experience, a critical element of effective benchmarks is that they have widespread legitimacy, which can be conferred through official endorsement by experts or policy makers.

However, using data to set benchmarks has the potential to increase their legitimacy and utility in a several ways.

The EGRA toolkit (developed for USAID by RTI, RTI International, 2015) recommends to benchmark fluency against comprehension and to assess current achievement levels in order for benchmarks to be “ambitious, but realistic and achievable” (p.132). Data can also be used to assess the predictive validity of benchmarks. Two steps are recommended in the benchmark setting process. First, discuss an acceptable level of comprehension with stakeholders. Second, identify a range of fluency associated with this level of comprehension.

Several countries have multiple benchmarks at different levels creating three or even four categories of readers. Composite benchmarks – where scores on several reading skills are combined to create a benchmark – lack validity and are not recommended.

Most countries set different benchmarks for each grade. It is important to ensure such benchmarks are created such that students moving up a grade are not reclassified with a lower level of reading proficiency. Other countries have chosen to set the same benchmark for Grades 1 and 2 with different targets for each grade.

Data from 35 language-specific benchmarks in 20 countries show that the majority of benchmarks set are in the range of 40–50 cwpm. The proportion of students reaching the benchmark in Asian countries has been higher (median 29%) than in Africa and the Middle East (median 5%). Benchmarks are more useful in tracking system improvements when a reasonable proportion of students achieve the benchmark, as is the case in many Asian countries. Countries with the highest level of achievement may consider assessing the predictive relationship between fluency and national assessments in order to set benchmarks at a higher level.

Room to Read has piloted use of a standard approach to benchmarks across countries. They use regression equations to model the relationship between comprehension and fluency.

These analyses point to preliminary conclusions about the factors required for accurate benchmarks, namely: a strong relationship between fluency and comprehension, use of reliable comprehension measures, good distribution of comprehension scores, and sample sizes of at least 150.

Recommendations

1. Sample sizes of less than 200 have typically produced unreliable benchmarks. In general, the larger the sample the more accurate the benchmark.
2. The sample should contain enough students around the level of the benchmark. Samples where very few students reach the comprehension threshold (e.g. 80%) or where almost all students reach the threshold are not useful for benchmarking.
3. Reliable and valid comprehension measures should be used. When comprehension is measured through reading a passage, students should be given enough time to read to the end of the passage. Consider assessing comprehension independently from fluency – for example a separate comprehension subtest based on reading of a second passage. Assess the quality of comprehension measures by assessing the mean and standard deviation of scores on individual comprehension items, as well as the internal reliability of the measure (i.e., how scores on comprehension questions correlate). The comprehension measure should be piloted and assessed before conducting the benchmarking exercise.
4. Consider use of a more robust measure of comprehension in benchmarking exercises. For one example of such a test see Jukes et al (in press).
5. Use of at least two passages can improve the reliability of the benchmark. Benchmarks can vary with the difficulty level of a passage—both in terms of readability and comprehension. It may be particularly useful to select two passages that are equated for difficulty and are grade appropriate.
6. Benchmarks that are closer to the center of a distribution, like those in Asia, are likely to be more sensitive to improvement and, therefore, more useful for tracking progress. Under such circumstances, a simpler more efficient approach to benchmarking can reliably be taken, e.g., one benchmark for oral reading fluency could serve as an effective indicator. Countries in which students in early grades score at the upper end of achievement distribution may also consider the alternative approach of setting fluency benchmarks based on percentiles of the achievement distribution in a normative sample of students on a national, standardized test, rather than with reference to levels of comprehension.

1.4 Experience of Setting and Using Fluency Benchmarks

The process begins by deciding on the reading skills to benchmark. Most countries choose fluency. Other countries choose to benchmark lower order reading skills as well. The target grades for benchmarks should be agreed upon as well as the target languages and regions. Benchmarks should be language-specific. In any multilingual context, it is essential to ensure that appropriate regional representatives are involved in the benchmark setting activity for all relevant language groups.

The process of setting benchmarks should involve a wide group of regional and national stakeholders, particularly key decision makers from the national ministry of education to ensure buy-in for the benchmarks. Only with a strong sense of ownership, understanding, and acceptance will benchmarks have an opportunity to be institutionalized. It is important to make use of the extensive knowledge of stakeholders to settle on benchmarks that are not

only scientifically (i.e., supported by data) defensible but also politically feasible, aligned with learning expectations, and ultimately “make sense”.

External organization with experience of benchmark setting can play an important role. It is essential that they can provide technical expertise and assistance without steering the process toward a pre-determined outcome.

When setting benchmarks and targets, it is necessary to strike the appropriate balance between science – data analysis that is evident to stakeholders - and art – guiding dialogue to broker agreement among different viewpoints.

After benchmarks have been set, existing data should be used to determine the current (or baseline) percentage of students meeting each benchmark. Ideally, these data should be derived from intervention studies to improve reading skills. Alternatively, the use of performance data over several years or, failing that, data from multiple grades in one year, can be used to estimate expected improvement from one grade (or one year) to the next

Benchmarks should inform targets, not the other way around. Of vital importance is to help stakeholders resist the temptation to lower the benchmark so that a higher target can be more easily reached. Benchmarks should represent the actual desired level of skill acquisition, and targets should be realistic based on assumptions about how much improvement is achievable. If reading proficiency benchmarks are too ambitious, countries can choose to set benchmarks representing lower categories of reading achievement (for example for emergent or beginning reading).

It is important that benchmarks are adopted by countries. To promote institutionalization, conversations should start early to ensure that the government has a full understanding of how benchmarks will be set. The presence of a project using benchmarks over several years that models the utility of having benchmarks to track progress increases the likelihood of them being officially adopted.

Recommendations

1. The aims and scope of the benchmarking activity must be clearly defined.
2. Relevant data must be obtained to address the pre-defined aims.
3. Benchmarks should be set in a participatory workshop that involves representation from a range of stakeholder groups.
4. Short- and long-term targets (based on the newly defined benchmarks) should be agreed upon.
5. Benchmarks and targets should be disseminated to obtain wide-ranging approval and institutionalization.
6. Projects can help ministries officially institutionalize benchmarks by modeling how using those benchmarks allows subsequent rounds of early grade reading assessments to show progress in terms of the percentages and numbers of children achieving levels of reading deemed to be proficient.
7. As additional data become available, projects can also help ministries revisit and reevaluate their benchmarks to ensure their reliability.

Conclusion

A last point concerns the increased importance of setting benchmarks now that the Sustainable Development Goal (SDG) for education includes an indicator (4.1.1) on the proportion of children achieving at least a minimum proficiency level in reading (and math). The global dialogue about SDG indicator 4.1.1 has recognized the linguistic and orthographic differences across languages (as discussed in this paper) and recognizes the different development levels of each country's education system. Therefore, it is accepted that each country determines its own definition of "minimum proficiency." The analyses and processes described in this paper for setting benchmarks are intended to help countries do exactly that. With defined benchmarks, countries can then measure and report on their progress in meeting the education-related sustainable development goals.

2 Introduction

2.1 Purpose and Outline of Report

Attending school and becoming literate is a foundation for success in life. Literate societies have higher life expectancy rates, lower crime, less teen pregnancy, and lower infant mortality (Burchfield, Hau, Baral, & Rocha, 2002; Sen, 1997; McMahon, 2000; McMahon, 2002; Wolfe and Haverman, 2002; United Nations Educational, Scientific and Cultural Organization [UNESCO], 2005). Literate communities are also less violent and more stable (Yanagizawa-Drott, 2012). To monitor and predict the extent to which the benefits of education will accrue to an increasing proportion of society, countries (and the United States Agency for International Development [USAID]) need to be able to monitor whether students meet a standard for reading proficiency during their early primary education. An education system that increasingly produces proficient readers indicates that a country is increasing the likelihood that its future citizens will lead productive and engaged lives. Students learning to read early also will predicate a decreased likelihood that citizens will lead lives that are counterproductive to a country's goals.

Beginning in the early 2000s, increasingly available data on learning outcomes, through international assessments, such as Trends in International Mathematics and Science Study, Program for International Student Assessment, and Progress in International Reading Literacy Study, showed how far behind students in developing countries were when compared to their peers in more affluent countries. However, one issue with these assessments was that they relied on students having a basic reading ability for administration; therefore, they were unable to evaluate students at lower levels of skill development, which was a necessity in lower-performing countries. This paved the way for the 2007 introduction of the individually, orally administered Early Grade Reading Assessment (EGRA). EGRA was designed to measure not only whether children could read and comprehend grade level text, but also measure how well children acquired basic literacy subskills, such as phonemic awareness, letter sound recognition, decoding, and familiar word recognition. As such, it offered a way to probe beneath the floor levels of performance that international comparative assessments showed for many developing countries. Because EGRA was open sourced and readily adaptable to the relevant language of instruction, it spread rapidly to more than 70 countries and was translated into 120 different languages in less than 10 years.

The deployment of EGRA began demonstrating that children in some developing countries were passing through several years of primary school without achieving the most fundamental of educational outcomes—learning to read. If countries were concerned with improving the performance of their students, then assuring acquisition of this fundamental skill was paramount. It is self-evident that students who do not learn to read will not be able to learn other subjects and will have a greatly reduced likelihood of being able to successfully complete primary school. Moreover, students who do not acquire strong foundational skills essential to learning to read in the first years of schooling are very likely to fall further behind their peers (RTI International, 2015). Benchmarks were introduced to simplify EGRA results into a single indicator against which countries could measure their children's reading progress. The following sections explain the rationale, processes, and considerations necessary for setting accurate and actionable benchmarks and targets (where benchmarks represent a desired level of performance on a reading task and targets represent the percent of students intended to be reaching the performance level in the future).

Overall, this report reviews experiences using EGRAs to set fluency benchmarks for a range of education systems with the aim of informing good practice in USAID-supported projects in Asia. The report draws on the experiences of multiple organizations that support national

governments to define and use benchmarks for oral reading fluency. Our approach to the report acknowledges multiple aims of fluency benchmarking, as noted below.

1. Country purpose: Track progress in reading within a national education system and provide data to inform efforts to improve education quality.
2. Agency purpose: Help USAID monitor the progress of projects and countries working to reach specific children-reading goals.
3. Global purpose: Measure progress in reading for all; provide a basis for global advocacy; and provide a method of assessing Indicator 4.1.1 of the Sustainable Development Goals: The proportion of children and young people (a) in grades 2 and 3, (b) at the end of primary, and (c) at the end of lower secondary achieving at least a minimum proficiency level in (1) reading and (2) mathematics, by sex.

To address these purposes, an understanding of the use and function of EGRA is required (as outlined in the following section). The global use of benchmarking data in the first and second purposes mentioned above relies on the technical work of defining viable and reliable benchmarks. That work is informed by an understanding of the science of reading development (Section 2) and by particular approaches to data collection and analysis (Section 3). The use of benchmarks within national education systems also depends on the process by which benchmarks are set (Section 4). Finally, in Section 5, we summarize the implications of our findings for use of benchmarks in Asia.

Our recommendations for the practice of benchmarking in Asia are based on review of data and experiences from Asia and from other continents. Where possible, we have focused our report on Asian examples. However, we also wanted to learn from a wealth of data and experience from Africa and the Middle East. We carefully considered the applicability of these experiences to Asia by considering similarities and differences in context along a number of dimensions, including linguistic factors, levels of academic achievement, and systems and processes of government. This enabled us to identify where experiences in Africa and the Middle East may be directly applicable to Asian countries and where approaches used in those regions may need to be changed to be applicable in some Asian countries.

3 Development and Measurement of Reading

To be able to set appropriate and usable early grade reading benchmarks, it is first necessary to understand what reading proficiency is and how to measure it. In this section, we review the science of reading development, with a focus on how it varies across languages and what factors may be relevant for Asian languages.

3.1 Stages of Reading Development

The development of reading proficiency has often been construed as progressing through a series of discrete but interrelated stages. Jeanne Chall's (1983) five-step model, which grew out of her seminal work in the late 1960s and continues to influence research and practice (Carnine, Sibert, Kame'enui, & Tarver, 2014), comprises a series of continuous and overlapping stages, each of which depends on success in prior stages. This model begins with initial reading, progresses through greater awareness of the alphabetic principle and decoding, and culminates in an increasingly sophisticated ability to read a broad range of texts and greater levels of complexity. Other models of reading development focus more intentionally on initial logographic processing or identifying meaning based on the visual characteristics of a letter or word (Bastien-Toniazzo & Jullien, 2001; Frith, 1985). For example, Frith's three-stage model describes literacy acquisition as progressing through a logographic stage of rote learning words based on their visual characteristics, an alphabet or phonological stage during which letter sounds and word elements are learned and words are

decoded phonetically. This results in an orthographic or sight word reading stage, when reading is more fluent and automatic, with more attention focused on comprehension (Frith, 1985).

Increasingly, however, reading development is being conceptualized less as a sequence of stages and more as an integrated series of skills and knowledge that develop concurrently, iteratively, and synergistically (Bulat, et al., 2017). Jackson and Coltheart (2001), for example, describe reading as an information-processing system that, itself, changes as an emerging reader gains proficiency in translating print to pronunciation, recognizing strings of letters, and accessing meanings of single words.

Whether sequential or concurrent, there is much commonality in the basic skills required to obtain reading proficiency. ***Table 1*** shows the individual skills that children need to acquire to be a proficient reader and how EGRA was designed to mirror this skills sequence. A more detailed discussion of each possible EGRA subtest is available in the Second Edition of the EGRA Toolkit (RTI International, 2015).

Table 1: Relationship between EGRA subtasks and early literacy skills

Being a proficient reader requires that students	Corresponding EGRA Subtests
Understand the language of the text	**Listening comprehension** is used as a proxy for student oral language development
Know the individual sounds that make up words	**Phonological awareness** measures students' ability to discriminate the individual sounds heard at the beginning, middle, and end of words
Know letters and letter sounds	The **letter names and sounds** subtests are used to measure how automatically children can recognize letters and produce the sounds those letters make
Decode unfamiliar words	**Non-word reading** evaluates whether students can combine letter sounds to read words that they do not automatically recognize
Recognize and read familiar words	**Familiar word reading** tests how automatically (i.e., fluently) children can read commonly used words taken from the in-school vocabulary for their grade level
Read text well enough to understand it	**Oral reading of a text passage** reliably gauges whether students are processing text fluently enough to move through it at a pace that does not hinder comprehension
	Reading comprehension is also assessed directly through questions about the text children read orally

Among this sequence of reading skills, there is a recognition of the critical relationship between fluency and comprehension and the need to achieve both automaticity and accuracy in reading individual words, followed by words strung together, sentences, and longer pieces of connected text (i.e., a reading passage). Until this word-level reading is mastered, most effort is exerted in simply decoding words, and insufficient attention can be given to making meaning of these words (Rasinski, 2011). One goal in reading instruction, therefore, is to build speed and accuracy of word reading so that cognition can focus on comprehension (Samuels, 2002).

3.2 Role of Reading Accuracy and Speed

Looking into the component mechanics that make up fluency gives a more detailed understanding of its relationship with comprehension. There is a body of research dating

back to the 1970s that shows accuracy is an important predictor of comprehension in the early stages of reading but that reading speed takes over as a better predictor once a child surpasses basic reading ability. Reaching a point of automaticity in reading words (i.e., reading words accurately and immediately, without having to sound them out) is important to facilitate comprehension. However, even after automaticity is achieved, reading speed continues to increase as individuals continue to gain proficiency (Stanovich, 2000). In fact, findings from studies such as those conducted by Hogaboam and Perfetti (1975) suggest that once a child reaches basic levels of reading fluency, it is reading speed rather than automaticity that fuels ongoing growth in reading proficiency (Stanovich, 2000). Mason (1980) also presented evidence suggesting that individual differences in reading proficiency can be, in part, explained by differences in perceptual processing speed.

Although there is a dearth of literature directly addressing fluency benchmarking, numerous studies have demonstrated a strong concurrent and predictive relationship between oral reading fluency and reading comprehension (Daane, Campbell, Grigg, Goodman, & Oranje, 2005; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Hudson, Pullen, Lane, & Torgesen, 2009; Kim, Petscher, Schatschneider, & Foorman, 2010; Kuhn, Schwanenflugel, & Meisinger, 2010; Kuhn & Stahl, 2003). This predictive relationship develops with reading proficiency (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003). At early stages of reading development, fluency depends on children's ability to read words proficiently (Kim, Wagner, & Lopez, 2012; Samuels, 2006; Wolf & Katzir-Cohen, 2001). After children have developed sufficient word reading proficiency, comprehension skills (both listening and reading) become important determinants of fluency (Kim, Park, & Wagner, 2014; Kim, 2015; Kim & Wagner, 2015; Petscher & Kim, 2011).

One longitudinal study in the United States (Good, Simmons, & Kame'enui, 2001) has examined the predictive validity specifically of fluency benchmarks. Results showed that a first-grade fluency benchmark had good predictive validity for achieving the second-grade benchmark. Of students who attained the first-grade benchmark goal (>= 40 cwpm in grade-level material) in this study, 97% attained the second-grade benchmark goal (>=90 cwpm). Of those categorized as needing instructional support (<10 cwpm) in grade 1, none attained the second-grade goal. The third-grade benchmark had good predictive validity for a standardized statewide high-stakes assessment. Of those achieving the third-grade benchmark goal, 96% were rated as "meets expectations" in the high-stakes assessment. Only 28% of those identified as having instructional needs (< 70 cwpm) achieved the "meetings expectations" rating in the high-stakes assessment.

Although it is clear that the relationship between fluency and comprehension is essential for proficient reading, there are three important implications for benchmarking work with Asian languages: (1) development of reading fluency in non-alphabetic languages may take longer to achieve than in alphabetic languages, (2) the strength of the relationship between fluency and comprehension may vary due (in part) to the depth of a language's orthography, and (3) there are some challenges in applying the same method of assessing oral reading fluency across languages.

3.3 Linguistic Differences in the Development of Reading Fluency

Due to their visual complexity and the amount of information that can be conveyed by a single character, development of reading fluency in non-alphabetic languages may take longer to achieve than it would in alphabetic languages (Liu, Chen, Liu, & Fu, 2012; Nakamura & de Hoop, 2014). In addition, the number of graphemes that exist across languages can further impact the rate of reading fluency acquisition (Chang, Plaut, & Perfetti, 2015). In fact, the growing body of research estimates that in alphabetic orthographies, which have on average 20–30 graphemes, and especially shallow orthographies in which there are clear, one to one relationships between letters and their sounds, many children can master all graphemes after one year of formal instruction. In alphasyllabic orthographies, which can have on average 400 graphemes, children can

require 3 to 4 years of formal instruction to achieve fluency. Whereas in logographic orthographies, such as Chinese, which can have more than 3,000 graphemes, it can take six or more years of formal instruction to achieve fluency (Chang et al., 2015). In a comparison of reading development across language with deep (e.g., English and French) and shallow (e.g., Finnish and German) orthographies, Ellis et al., (2004) found that children learning to read in shallow orthographies gained reading fluency more than twice as fast as those reading the deep orthographic language of English.

3.4 Linguistic Differences in the Fluency-Comprehension Relationship

In opaque (or deep) orthographies, such as English, children's ability to read a passage fluently requires a combination of word reading skills and comprehension of the context in which the word appears (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003; Wolf & Katzir-Cohen, 2001; Kim, Wagner, & Foster, 2011). For example, the word "bow" in English can be pronounced in one of two ways and understanding the context in which it appears is essential for its correct pronunciation. In contrast, in a transparent orthography, words can be pronounced correctly without understanding their context. Therefore, we may expect the fluency-comprehension relationship to be stronger in opaque vs transparent orthographies.

Other aspects of a language's orthography can affect the fluency-comprehension relationship. Abadzi (2011) notes that a "higher reading speed may be necessary to comprehend what is read for orthographies that, for example, omit vowels and force readers to keep alternative pronunciations of words in working memory to make sense of a sentence (e.g., unvoweled Arabic, Hebrew, Farsi, Pashto, and Urdu). On the other hand, it may be possible to read more slowly and comprehend tonal languages, where a morpheme carries two bits of information (e.g., marked tones and short words in Lao, Vietnamese, and Thai)." The way in which a language marks, or does not mark, word and sentence boundaries can also impact the speed of reading fluency and, consequently, comprehension. Many Asian languages, such as Chinese, Lao, Thai, and Khmer (an alphasyllabary), do not provide word boundaries, which introduces complexities in reading connected text. Rather than following a relatively linear process of either encoding letter sounds to identify the word or identifying the word by sight, reading text without word boundaries is much more iterative (see Winskel, 2014, for instance). According to Shen and Jiang (2013), reading connected text in such languages is an interactive process, and inaccurate lexical access may cause difficulty in text comprehension, which then notifies the reader that he/she should revisit the word segmentation process. As a result of this iterative process, we may expect a stronger relationship between fluency and comprehension in languages where word boundaries are not clear (e.g. Chinese, Lao, Thai, and Khmer). Readers need to understand the meaning of the text to infer the position of word boundaries and, thus, be able to read the text fluently.

The above discussion outlines the orthographic features that may influence the relationship between comprehension and fluency. What is the evidence of this relationship across languages? As would be expected for a language with an opaque orthography, in English the relationship between reading fluency and comprehension has been broadly established (for example, at a correlation of r=.91, Fuchs, Fuchs, Hops, & Jenkins, 2001). The relationship is also strong for non-English speakers learning English as a second language (Pretorius & Spaull, 2016), although the strength of that relationship can vary for a child's first, second, or third language (Piper, Schroeder, & Trudell, 2016). Research has also established a moderate-to-strong relationship between comprehension and fluency in a range of non-English languages, e.g., Turkish (Basaran, 2013), and transparent bantu languages such as Kiswahili and Gikuyu (Piper, Schroeder, & Trudell, 2016). A far smaller body of research exists, however, for alphabsyllabic languages (such as Korean, modern Lao, and Khmer script) and logographic languages (such as Chinese and Japanese kanji)—and existing research has not yet provided consistent results. In Korean, Pae and Sevcik (2011) found a strong correlation between reading fluency and comprehension. Even in the logographic language of Chinese, researchers found moderate to moderate-high correlations

between comprehension and character-naming accuracy (r = .64) and character-naming speed (r =. 55) (Shen & Jiang, 2013). In the Indian alphasyllabic languages of Kannada or Telugu, however, Nakamura and de Hoop (2014) did not find the same strength of relationship. We conclude that fluency and comprehension are moderately to strongly related in all reported studies to date, but there is insufficient evidence to examine how linguistic differences impact this relationship.

3.5 Linguistic Differences in the Fluency Assessment

Oral reading fluency is typically assessed in terms of the number of words of a passage that are read correctly in one minute. Applying this metric can be challenging in languages where word boundaries are ambiguous. In Chinese, there is ambiguity in whether certain character strings should be read as one word or multiple words (such as “boy” versus “male child”). Therefore, Shen and Jiang (2013) opted to measure fluency by counting word segmentation errors rather than accurate word segmentations. The Khmer language of Cambodia faces a similar problem. For example, the Khmer word for “shoe” translates as “leather foot” which may be considered two words or a single compound word. To resolve potential ambiguities in counting words in a passage, Room to Read (2016) convened a team of national experts to determine word boundaries in reading passages. A different approach was taken in Lao and Vietnamese as these languages are based around syllables that can be counted with less ambiguity than words. Consequently, Room to Read has assessed fluency in these languages expressed in syllables read per minute.

Despite the linguistic differences discussed above, what consistently emerges from the literature and experience is a recognition of the similarities in learning to read across language types. Even in alphasyllabic and logographic languages, foundational skills, such as phonological processing, orthographic processing, working memory, and vocabulary play a role (Cho & Chen, 1999; Pae & Sevcik, 2011; Shen & Jiang, 2013; Nakamura & de Hoop, 2014). Furthermore, it is evident that a universal theory of learning to read may well apply across all types of languages (Nag & Snowling, 2001).

3.6 Assessing Reading Proficiency: Why Focus on Fluency?

As previously noted, the EGRA provides a comprehensive view of a child’s reading skills (from alphabetic awareness to reading comprehension[1]). However, when monitoring progress and setting standards for an education system it is helpful to have a simple metric for and clear understanding of when reading proficiency has been achieved.

Many would argue that comprehension is the ultimate goal of any reading endeavor. It is possible to directly assess a reader’s ability to read connected text and fully comprehend what is read; however, it can be challenging to do so reliably in general (Sweet & Snow, 2003) and with the EGRA instrument in particular (Bartlett, Dowd, & Jonason, 2015). Measures of comprehension vary with a number of factors, e.g., a child’s familiarity with the subject matter of the text. This makes it difficult to set reliable benchmarks for comprehension using the EGRA or similar assessments.

An alternative approach is to monitor children’s reading fluency. First, because fluency is an important skill in its own right—some children can comprehend text but lack fluency (for example, of the children only able to read two sentences of the reading passage on the 2014 National EGRA in Indonesia, nearly half were able to answer both of the corresponding reading comprehension questions correctly). Second, it is straightforward to measure reading fluency rates. Third, fluency is a transparent measure. It is relatively easy to explain to parents and teachers how fluency is measured, why it is important, and what fluent

[1] While the five or six question reading comprehension portion of a typical EGRA is limited as an evaluation of reading comprehension, it does provide a measure of a student’s recall of basic information from the text she would have read (Dubeck & Gove, 2015).

reading looks and sounds like. Finally, and crucially, there is a high level of correlation between reading fluency and comprehension, as discussed above.

Oral reading fluency has regularly been used as a proxy measure of comprehension and overall reading ability over the past several decades (Abadzi, 2011). Both in the United States and internationally, 1-minute timed measures of oral reading fluency, often accompanied by comprehension questions, have become standard measures of reading proficiency and have been found to have high levels of validity and reliability (see, for example, Stage & Jacobson, 2001). For the most part, these measures are used to identify and track readers' levels of performance within a given language; increasingly, however, research is exploring their efficacy in describing reading proficiency trends across languages. When doing so, it is critical to understand the languages involved, because as described earlier, language characteristics can impact the rate at which fluency is achieved.

For example, some Asian languages can be written in two forms—with or without diacritics. Diacritics are symbols indicating vowel sounds that can be attached to consonants and writing with diacritics is a shorter form of writing. Diacritics exist in a several Asian languages, such as Arabic and Hindi, where they are known as Matras. One implication for assessing fluency is that fluency rates will differ in the two script forms. Also, in some languages, diacritics are taught later in the school curriculum. Therefore, assessing students on text with diacritics would not be grade-appropriate in some of the early primary school years.

Lastly, the EGRA comprehension measure is limited in that it typically only contains five items. This can make analyses (including basic psychometrics for reliability/validity, as well as more complex approaches such as equating across assessments) very complicated. Therefore, oral reading fluency has generally been used as a proxy for comprehension in the majority of EGRA work. Recently, there have been efforts to develop other options for evaluating comprehension as part of an EGRA[2], but nearly all work to date has relied on the five-item reading comprehension subtask.

Based on the above considerations, when establishing benchmarks for reading proficiency, as measured by oral reading fluency rates, it is important to note that a benchmark established for one language is not necessarily relevant to other languages, given differences between languages' linguistic and orthographic features (e.g., word length and complexity) and depth of orthography (Graham and van Ginkel, 2014). When assessing and comparing emergent readers, these differences are of particular concern (Abadzi, 2011).

3.7 From Fluency to Benchmarks

The above discussion establishes oral reading fluency as a plausible metric to track reading proficiency. This report examines the experience of deriving benchmarks from fluency measures. While it is possible to track population-level improvements in reading proficiency in other ways (e.g., by calculating mean fluency rates), there are a number of advantages to converting fluency scores into benchmarks.

First, the process of setting benchmarks allows education systems to articulate their definition of reading proficiency. Second, the use of benchmarks communicates this definition of reading proficiency to others and can provide a target for teachers and students to aim for. Third, this approach allows a direct assessment of how many students are reading proficiently, rather than, for example, a measure of improvement in mean fluency, which may not be as informative. Counts (or percentages) of children are more readily interpreted, especially by non-experts. Counts can also be summed across contexts, for example to assess progress against the USAID All-Children-Reading targets.

[2] See for example Jukes et al. (in press) for discussion of alternative measures of comprehension.

As with any way of summarizing data, information is lost by converting raw fluency scores to binary indicators of whether a benchmark is achieved. In particular, where student achievement is low, progress can be made in helping students learn basic literacy skills without this improvement being evident in reported statistics related to the number of children reaching the fluency benchmark. Methods for setting fluency benchmarks and for tracking progress, including those which attempt to account for and note progress below a proficiency standard, will be discussed in the next section.

Summary: Implications for Asian languages

Our understanding of reading fluency is derived from research conducted mostly with Western European alphabets. However, there is a great diversity of languages and writing systems in the developing world, and many of the non-alphabetic languages are in Asia. Based on our review of the science of reading development and reading assessment we can suggest various recommendations for benchmarking in Asia.

1. There are many similarities in the process of learning to read across languages. The same foundational processes are important in alphabetic, alphasyllabic, and logographic languages. The science of reading supports the assumption that EGRA assessments, and the use of fluency benchmarks, are appropriate to use across Asia.
2. Orthography influences the rate of language acquisition. For example, learning to read in Chinese is a much slower process than in alphabetic languages. Consequently, progress in achieving reading proficiency should not be compared across languages and more modest targets should be set in more complex languages.
3. Fluency is typically measured in a count of words read per minute. Counting words in logographic or alphasyllabic languages can be challenging because of ambiguities in whether some compound words are counted as one or two words (e.g., “leather foot” is “shoe” in Khmer). Some approaches to this issue involve counting characters or syllables rather than words, focusing on errors in word segmentation rather than accurate word segmentation, or convening experts to adjudicate on the count of words.
4. Tentative conclusions from research to date are that oral reading fluency is a good proxy for comprehension across languages. However, there are theoretical reasons why the strength of the relationship between fluency and comprehension may vary with orthography. When working in Asian languages that have not been the subject of much empirical research, benchmarking exercises could and should seek to test the hypothesis that fluency is a good proxy for comprehension in those languages.
5. Some writing systems have different versions, with or without the use of diacritics. One implication for assessing fluency is that fluency rates differ in the two forms of the script. And the version to be used when assessing students will depend on when the curriculum introduces students to the use of diacritics.

4 Fluency Benchmarking Data

4.1 Using Data to Set Fluency Benchmarks

To understand the role of data in setting fluency benchmarks, it helps to restate the goals of fluency benchmarks. They include:

- Defining proficiency in reading fluency in a given language and education system

- Providing a goal for students and educators
- Providing a means to track progress of an education system

It is possible that a fluency benchmark set without using data can fulfill these functions. A benchmark set, for example, using the judgement of a panel of experts may offer a credible definition of proficiency and a useful means to motivate and track educational progress. In our experience, a critical element of effective benchmarks is that they have widespread legitimacy, which can be conferred through official endorsement by experts or policy makers.

However, using data to set benchmarks has the potential to increase their legitimacy and utility in a several ways:

- Data are most commonly used to help define proficient reading. The EGRA toolkit (RTI International, 2015) recommends that reading proficiency be defined in terms of comprehension. Thus, data can be used to assess the level of reading fluency that is indicative of good comprehension. An alternative approach is to define proficiency solely in terms of fluency. This involves measuring fluency rates in a sample of proficient (typically older) readers.
- The EGRA toolkit recommends that benchmarks be "ambitious, but realistic and achievable" (p.132). Collection of data can give a better understanding of current achievement levels and the level at which an ambitious but realistic benchmark should be set.
- Benchmarks should have predictive validity (Dynamic Measurement Group, Inc., 2010). A student who achieves a benchmark goal should be more likely to achieve later reading outcomes. Data can be used to establish the predictive validity of benchmarks.

4.2 How Are Benchmarks Set Using EGRA Data?

The following steps are recommended in the EGRA toolkit (as prepared for USAID by RTI International, 2015, p. 133):

Step 1: Begin by discussing the level of reading comprehension that is acceptable as demonstrating full understanding of a given text. When EGRA data on reading comprehension are used, most countries have settled on 80% or higher (usually, 4 or more correct responses out of 5 questions) as the desirable level of comprehension. It would be worthwhile to explore how other measures of comprehension could further inform the benchmarking process.

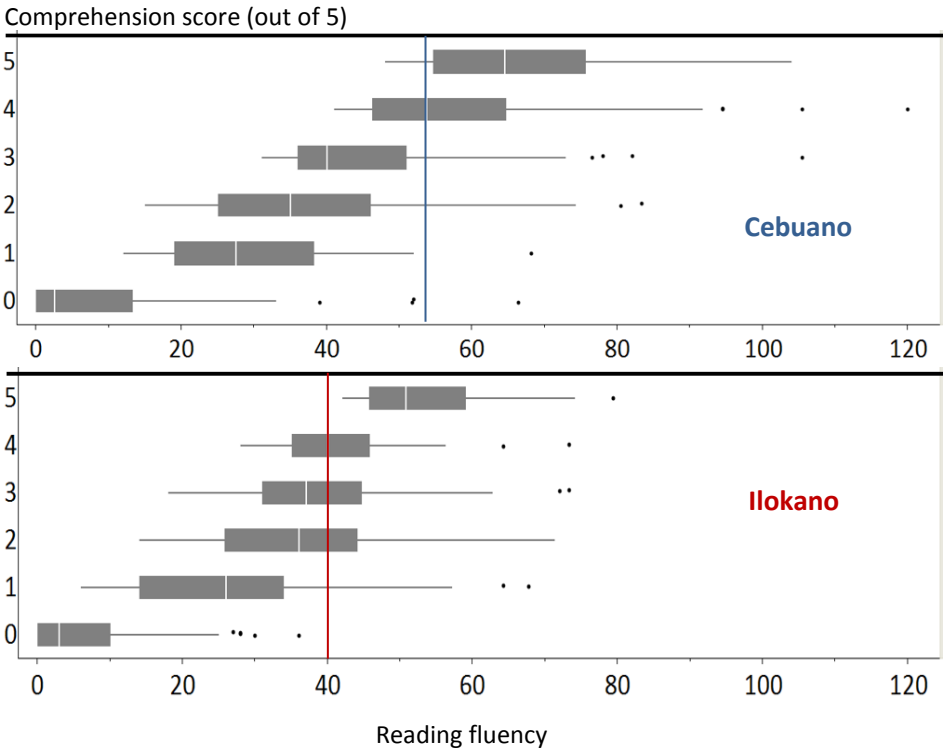
Step 2: Given a reading comprehension benchmark, EGRA data are used to show the range of oral reading fluency (ORF) scores—measured in correct words per minute (cwpm)—obtained by students able to achieve the desired level of comprehension. Discussion then is needed to determine the value within that range that is put forward as the benchmark. Alternatively, a range can indicate the levels of skill development that are acceptable as "proficient" or meeting a grade-level standard (for example, 40 to 50 cwpm).

Most countries choose to present data on the relationship between fluency and comprehension to stakeholders to inform discussion around benchmark setting (see Section 4). ***Figure 1*** shows how this relationship was represented for two of the languages in the Philippines. The box-and-whisker plots summarize grade 2 student performance in Cebuano and Ilokano. Each row plotted vertically (along the y-axis) for each language corresponds to a level of reading comprehension, ranging from 0 to 5 questions answered correctly (out of 5). The horizontal axis shows reading fluency. Each box is based on the reading fluency scores at the 25^{th}, 50^{th}, and 75^{th} percentile for each level of reading comprehension. In every country where Education Data for Decision Making (EdData)II supported work on setting benchmarks, 4 correct answers out of 5, or 80% comprehension, was agreed to by

stakeholders as the standard for students demonstrating an acceptable level of comprehension.

The analysis that relates comprehension to fluency (as depicted in Figure 1) is less than exact because of the limitations of the comprehension measure usually included in EGRA. However, the point is not to present the relationship as determinant of the exact rate of fluency at which comprehension at the desired level is assured. The point is to illustrate to the people considering where to set the benchmark that the desired level of comprehension (as measured by EGRA) is associated with a range of reading fluency.

Figure 1: Distributions of reading fluency versus comprehension in two Philippine languages in grade 2



For example, the 25th percentile of students who answered 4 out 5 questions correctly in Cebuano had a fluency rate of 47 cwpm (the left edge of the box). The 50th percentile (the blue line) had a rate of 54 cwpm, while the 75th (right edge of the box) had a rate of 65 cwpm. The figure shows that for Ilokano, students achieving 4 out of 5 correct in comprehension were reading at fluency levels lower than their peers reading in Cebuano: the 50th percentile of students scoring 4 out of 5 was 40 cwpm in Ilokano.

These differences in the ranges of fluency levels associated with at least 80% comprehension were discussed, and participants in the benchmark setting workshop offered explanations based on the differences in the linguistic characteristics of the two languages. For example, Philippine Department of Education experts explained that Ilokano is an agglutinating language in which “words” can contain several units of meaning (e.g., subject, predicate, and object). Accordingly, students can glean more meaning from text, even though they may, technically, be reading fewer words correctly per minute as measured by EGRA. A lower benchmark for proficiency in Ilokano would, therefore, account for this.

As mentioned above, it is important to also point out that for every language there is a range of fluency scores (the boxes in the above plots) that correspond to the desired level of comprehension. Therefore, discussion as to what level of fluency should be set as a benchmark (i.e., one that shows that students are reading with comprehension) is required.

In almost all countries, this leads to vigorous discussion among stakeholders as to what the benchmark should be and, ultimately, results in greater ownership of the benchmark they finally decide to put forward.

As discussed above there are limitations to this method of identifying fluency benchmarks. Most notably, the weakness of the comprehension measure included in a typical EGRA limits the precision of any benchmark. However, as argued above, the point is not to calculate a precise benchmark but to engage stakeholders in considering the range of possible values that can be considered indicative of a proficient level of reading. Additional limitations to this approach, a technique for testing the reliability of a benchmark, and an alternative method for calculating benchmarks are discussed in Section 3.8 below and extensively in Jukes et al. (in press).

4.3 Setting Multiple Benchmarks

In addition to determining a language-specific benchmark for reading proficiency, some countries and some USAID programs were interested in tracking performance across levels of reading skill development. Several countries took different approaches to defining levels of reading ability. All cases shared one common classification, non-readers defined as students scoring zero on the oral reading fluency task.

In Ethiopia, data were used to establish grade-specific cutoff points (distinct for each language) for three classifications of reading ability above zero:

- Reading fluently with full comprehension
- Reading with increasing fluency and comprehension
- Reading slowly and with limited comprehension

In Pakistan, three levels of reading ability were also defined for each grade level for the two languages used as media of instruction in the regions where the USAID project was working. Based on their level of reading fluency, students would fall into three categories as shown in ***Table 2***. Oral reading fluency cutoff points for each category (measured in cwpm) are also shown for the two languages.

Table 2: Reading levels in Pakistan

	Grade 1		Grade 2		Grade 3	
	Urdu	**Sindhi**	**Urdu**	**Sindhi**	**Urdu**	**Sindhi**
Does not meet expectation	< 30	< 30	< 60	< 50	< 70	< 60
Meets expectation	30 to 60	30 to 50	60 to 90	50 to 80	70 to 100	60 to 90
Exceeds expectation	> 60	> 50	> 90	> 80	> 100	> 90

Zambia employed a similar approach to Pakistan, defining two classifications other than non-reading: (1) emergent readers and (2) readers. Students achieving ORF of at least 20 cwpm and at least 40% comprehension (2 out of 5 questions correct) were defined as emergent readers. Those reaching at least 45 cwpm and 80% comprehension were defined as readers.

The **Kenya** Tusome Program set benchmarks at two levels—one corresponding to non-zero comprehension (emergent reader), and one at 75% comprehension (fluent reader). The use of a lower benchmark helped to quickly demonstrate progress and get buy-in for the benchmarks. The two benchmarks in Kiswahili are 17 and 45 cwpm and in English are 30 and 65 cwpm.

Tajikistan and the **Kyrgyz Republic** produced four categories of proficiency for a range of reading skills for grades 1, 2, and 4. ***Table 3*** shows how reading fluency was defined in these four categories in Tajikistan. The definitions apply to both Tajik and Russian.

Table 3: Definition of multiple fluency benchmarks in Tajikistan

Grade	Non satisfactory	Satisfactory	Good	Excellent
1	1–24 words	25–29 words	30–34 words	35+ words
2	1–39 words	40–44 words	45–49 words	50+ words
4	1–79 words	80–84 words	85–89 words	90+ words

The **Kyrgyz Republic** similarly used four categories of reading skill proficiencies

In each of these cases, benchmarks defined in relation to different levels of reading ability allow interested parties to track improvement below proficient reading. For example, if in Tajikistan, many students are not reaching the benchmark for reading fluently with comprehension, it is possible to monitor how cohorts are progressing through the other levels of performance and if the distribution of students across those levels appears to be improving over time.

4.4 Composite Benchmarks

An initial approach taken by the Quality Reading Project (QRP) in Tajikistan and the Kyrgyz Republic was to set a composite benchmark based on three sub-tests. There are two ways to construct a composite benchmark, using either a *compensatory* or *conjunctive* model. In a compensatory model, a student can perform poorly in one sub-test and still achieve the benchmark by performing strongly in another test. The weakness of a compensatory approach is that students may score poorly in reading fluency, but score highly in a lower order skill like phonemic awareness. Being good at the latter does not compensate for performing poorly at the desired higher order skill of reading fluently. Therefore, the compensatory approach would overstate what such a student is actually able to do.

In a conjunctive model, minimum performance requirements are set for each subtest, and a student needs to meet all requirements to achieve benchmark performance. The model reported for the baseline EGRA in both Tajikistan and the Kyrgyz Republic was a conjunctive model, in which the benchmark could only be achieved by performing well in familiar word reading, unfamiliar word reading, and ORF (of connected text). The results of this conjunctive benchmarking exercise in the Kyrgyz Republic are presented in ***Table 4*** below.

Table 4: Percentage of students passing individual and composite benchmarks in Russian in Kyrgyz Republic.

	Familiar Words	Unfamiliar Words	Fluency	Composite
Grade 1	46.8%	17.4%	36.6%	16.2%
Grade 2	58.6%	11.9%	44.5%	10.0%
Grade 4	23.3%	1.9%	32.4%	1.5%

As can be seen in ***Table 4***, the percentage of students meeting the composite benchmark in each grade masks considerable variation across grades in the areas of strength and weakness. The additional disadvantage of this approach is that achievement of the composite benchmark can be dominated by low achievement on one challenging subtest. In this example, achievement of the composite benchmark is determined almost entirely by proficiency in the unfamiliar words subtask. Clearly, this approach loses a lot of information

about the performance of students in the familiar words and fluency subtests. For this reason, the composite benchmark approach was wisely dropped in the QRP. Furthermore, this approach is not endorsed by USAID.

4.5 Benchmarks by Grade

Different countries have taken different approaches to setting benchmarks by grade. Some have set a different benchmark for each grade (see ***Table 2*** for Pakistan, above). In such cases, data can be used to calibrate benchmarks so that a different level of performance is set as the benchmark for each grade, but done so in a way that ensures that students moving up a grade are not reclassified with a lower level of reading proficiency as an artifact of a benchmark set too high for that subsequent grade (Zieky & Perie, 2006).

Other countries have maintained the same benchmark across multiple grades. For example, the Kenya Tusome Program has the same benchmarks for grade 1 and 2. In such cases, countries may choose to set different targets for the number of children reaching the benchmark in each grade. We discuss the issue of balancing targets and benchmarks in Section 4.

4.6 Results of Fluency Benchmarks Exercises in Asia, Africa, and the Middle East

Tables Table 5 ***andTable*** 6 show the benchmarks set for "proficient" reading (or reading with comprehension) in each language in either grade 2 or 3 in several countries.

Table 5: Benchmarks for reading proficiency in selected Asian countries

Country	Language	Fluency Benchmark	% Students Meeting the Benchmark
Cambodia[1]	Khmer	68 cwpm[a]	35%
Indonesia	Bahasa	59 cwpm	48%[b]
Kyrgyzstan[2]	Kyrgyz	40 cwpm	31%
	Russian	40 cwpm	49%
Pakistan	Urdu	60–90 cwpm	20%
	Sindhi	50–80 cwpm	24%
Papua New Guinea	English (WHP[c])	45 cwpm	1%
	English (Madang)	45 cwpm	8%
	English (NCD[d])	45 cwpm	8%
Philippines	Ilokano	40 cwpm	35%
	Hiligaynon	45 cwpm	34%
	Cebuano	42 cwpm	54%
	Maguindanaoan	40 cwpm	22%
Tajikistan[2]	Tajik	40 cwpm	35%
	Russian	40 cwpm	55%
Timor Leste	Tetum	45 cwpm	26%
Tonga[3]	Tongan (grade 2)	50 cwpm	15%
	Tongan (grade 3)	50 cwpm	34%
Vanuatu[3]	English (grade 2)	45 cwpm	6%

Country	Language	Fluency Benchmark	% Students Meeting the Benchmark
	English (grade 3)	45 cwpm	23%

1 - Data from Room to Read. Not a government-adopted benchmark.
2 - Data from AIR/QRP
3 - Data from PEARL program
a –At the time, Room to Read was using the mean correct words per minute for students with 80% comprehension as a way to approximate a benchmark.
b - Based on an unofficial benchmark of completing a 59-word passage
c - WHP = Western Highlands Province
d - NCD = National Capital District

Table 6: Benchmarks for reading proficiency in selected non-Asian countries

Country	Language	Fluency Benchmark [a]	% Students Meeting the Benchmark [a]
Egypt	Arabic	50 cwpm	11%
Ethiopia	Afaan Oromo	48 cwpm	5%
	Af Somali	50 cwpm	14%
	Amharic	50 cwpm	6%
	Hadiyyisa	40 cwpm	4%
	Sidamu Afoo	45 cwpm	1%
	Tigrinya	55 cwpm	<1%
	Wolayttatto	43 cwpm	8%
Ghana	Ghanaian languages [b]	40 cwpm	3%
	English	45 cwpm	7%
Jordan	Arabic	46 cwpm	3%
Kenya	Kiswahili	45 cwpm	n/a
	English	65 cwpm	34%[d]
Liberia	English	35–40 cwpm	4%
Malawi	Chichewa	40 cwpm	<1%
Tanzania	Kiswahili	50 cwpm	5%
West Bank	Arabic (with diacritics)	30 cwpm	18%
	Arabic (without diacritics)	35 cwpm	27%
Zambia	Zambian languages [c]	45 cwpm	1%

[a] Either grade 2 or grade 3, and from 2012 through 2014, depending on availability of EGRA data
[b] Akuapem-Twi, Asanti-Twi, Dagaare, Dagbani, Dangme, Ewe, Fante, Ga, Gonja, Kasem, and Nzema
[c] Chitonga, Cinyanja, Icibemba, Kiikaonde, Lunda, Luvale, and Silozi
[d] Baseline results

Tables Table 5 ***andTable 6*** illustrate how many countries adhered to the recommended practice of setting benchmarks separately by language, as was discussed in Section 2.

Of interest for Asian countries is where ***Table 6*** shows an example of how benchmarks were set separately for two versions of Arabic (with and without diacritics) in the West Bank. The West Bank data showed that reading with diacritics was more challenging for students, as evidenced by the fact that fewer students were able to achieve the lower benchmark level of reading fluency for Arabic with diacritics. As discussed in Section 2, a number of Asian languages use diacritics; therefore, the practice of setting separate benchmarks with and without diacritics could be applicable and useful. Comparing students' ability to reach benchmark levels of performance with and without diacritics could afford Asian countries with languages that use diacritics additional insight into how their use impacts students' reading performance.

Ghana and Zambia stand out as exceptions to the practice of setting separate benchmarks for each language. In these countries, a single benchmark for reading fluency was set for multiple languages. There was extremely low performance in reading across all the indigenous languages in the two countries, thus there were not enough students with high enough levels of fluency and comprehension to allow statistically valid estimates of a reasonable standard of proficiency. Therefore, a cross-language sample was used to set the benchmark.[3] This should not be an issue in most Asian countries where, in general, performance tends to be much better than what is seen when indigenous languages are introduced as media of instruction in African countries.

Two patterns are apparent across ***Tables Table*** 5 ***andTable*** 6. First, the majority of benchmarks set are in the range of 40–50 cwpm. There are a couple of notable exceptions to this. Kenya's benchmark for English was 65 cwpm. Pakistan adopted a benchmarking range which reached as high as 90 cwpm in Urdu and 80 cwpm in Sindhi. (Note that the high benchmarks presented for Cambodia and Indonesia results from statistical analyses presented in reports on reading fluency but have not been adopted by national governments in those countries).

Second, the proportion of students reaching the benchmark in Asian countries is higher than in Africa and the Middle East. As a rough guide, the median value among country- and region-level rates of passing the benchmark presented in ***Table 5*** is 29% and in ***Table 6*** is 5%. The tables are not comprehensive in their coverage of countries on the two continents, but they illustrate a trend which is consistent with other data (e.g., UNESCO, 2015).

4.7 Setting Benchmarks in High-Performing Education Systems

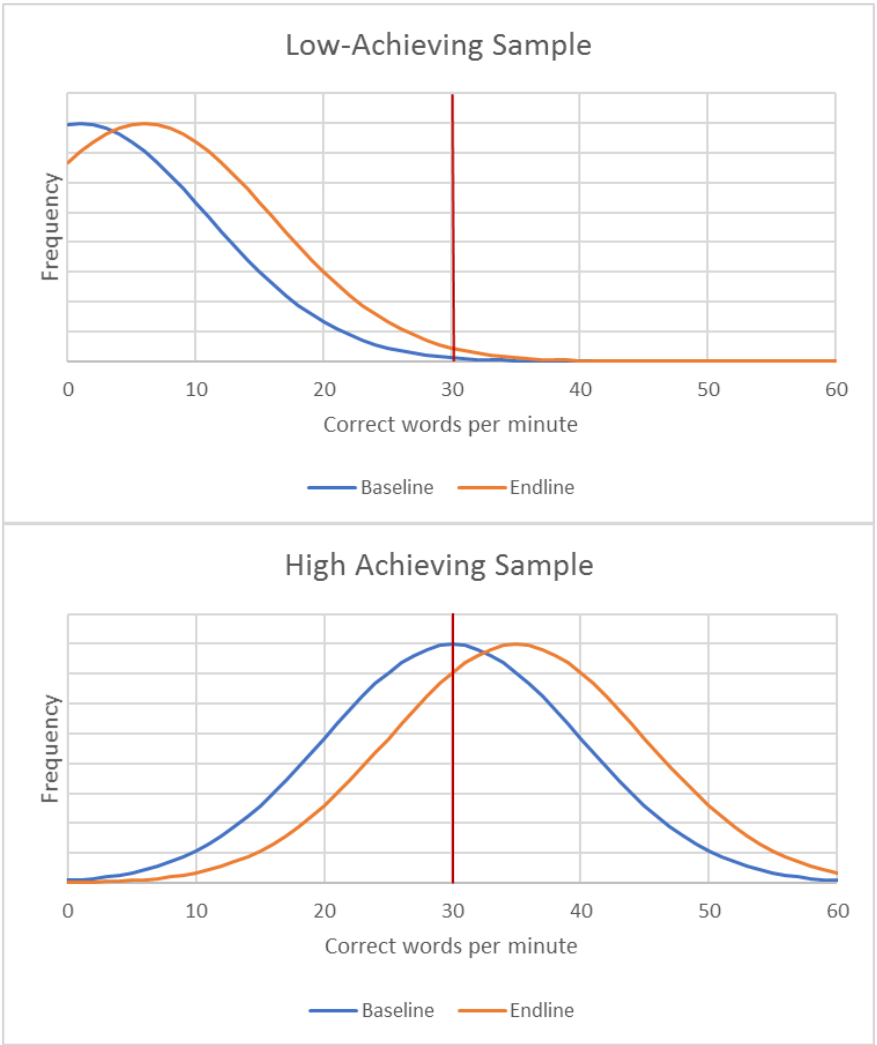
Comparison of ***Tables Table*** 5 ***andTable*** 6 suggests a trend in which Asian countries, compared to African countries, have more students in early grades performing above fluency benchmarks. Given the higher level of performance, is the benchmarking process used predominantly in lower-performing countries, going to be applicable to regions with higher-performing students?

In response to this question, we note that the percentage of children reaching the fluency benchmark in our sample of African countries clusters around 10% or less. With this low level of achievement, the approach to benchmarking in the region requires more careful use of data and often calls for setting multi-level benchmarks and benchmarks for a greater number of subskill areas (as ways to capture movement in student performance below a benchmark for reading fluently with comprehension, which few students will likely meet). The benchmarking process in higher-performing countries can be less complicated, and use of a single benchmark for reading fluently with comprehension could be an effective and efficient way to track progress. The percentage of students reaching the fluency benchmark in ***Table 5*** clusters around 30%. This implies that benchmarks in those countries are set nearer the center of the distribution of student performance. In such cases, the benchmark would be a more sensitive indicator of improvement. A small increase in mean performance could lead

[3] Benchmarks in those two countries should be revisited and language specific ones developed as additional data and more robust samples become available.

to a relatively large number of children (the bulge in the middle of the normal distribution) being promoted above the benchmark compared with countries where the benchmark is set in the tail of the distribution. The relationship between student achievement level and the sensitivity of benchmarks is illustrated in ***Figure 2***, which shows hypothetical distributions of oral reading fluency in low- and high-achieving countries, both using a fluency benchmark of 30 cwpm. In both countries, there is a modest mean improvement of 5 cwpm from baseline to end line. In the low-achieving country, this improvement leads to an increase of only 0.006 percentage points in the students passing the benchmarking, from 0.002% to 0.008%. In the high-achieving country the same mean improvement in ORF leads to an increase of 19 percentage points in the students passing the benchmark, from 50% to 69%.

Figure 2: Students reaching fluency benchmarks in low- and high-achieving samples



However, if some countries want to set more ambitious benchmarks, it is worth considering how other countries have similar, high benchmarks. Many school districts in the United States monitor reading achievement using the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) tests, developed by the Institute for Research and Learning Disabilities at the University of Minnesota in the 1970s and 80s. DIBELS use oral reading assessments akin to EGRA to evaluate individual students' development of early reading skills.[4] Extensive

[4] See http://www.literacyconnects.org/img/2013/03/DIBELS-and-what-they-measure.pdf

research since the 1980s, led primarily by the University of Oregon, has codified the reliability and validity of DIBELS, in particular showing how the test's benchmarks for different reading skill levels can reliably predict student performance in reading on standardized tests. In fact, the relationship between a student's performance in each skill area in grade 1 or 2, and his/her ability to meet grade level norms on a standardized test in grade 3 or 6 is what defines the benchmarks for each skill area.[5]

The University of Oregon, Center on Teaching and Learning, has supported the use of DIBELS across the United States since 2001. They provide benchmarks for grades K-6, for the beginning, middle, and end of the school year. The benchmarks represent the lowest core score at which students have a high likelihood of continued success. In addition, DIBELS provides a "cut score" that corresponds to the point at or below which students are at risk for not meeting grade-level expectations in the future.[6] For example, the oral reading fluency benchmark (for fluency alone, not in relation to any measure of comprehension) for English-speaking students in the US is set at 47 cwpm at the end of grade 1.[7] A student scoring at or above this benchmark would be associated with scoring in the 40th percentile or above on a nationally norm-referenced test. The cut score for being at risk at the end of first grade is 31 cwpm. Someone scoring at or below the cut score would be associated with scoring at or below the 20th percentile on a norm-referenced test. School districts have successfully been employing the DIBELS cut scores to identify and meet the needs of students who require additional, remedial instruction.

If high-achieving countries in Asia want to adopt a similar approach, work would be needed to align performance on an EGRA or EGRA-like assessment with performance on a national assessment, after which benchmarks could be based on the predictive relationship between reading fluency and points in the percentile distribution of national assessment results. This could bolster the applicability of the benchmarks as results would indicate not just whether a given level of fluency aligns with a desired level of comprehension, for example, in grade 2, but also whether the benchmark level of fluency in grade 2 predicts that the student is likely to be in the more desirable end of the distribution on a grade 3, 4, or 6 national exam.

4.8 New Approaches to Data Analysis for Fluency Benchmarking

As described above, data are typically used to provide a range of figures for benchmarking to guide dialogue with policymakers. Data can also be used to calculate a specific fluency benchmark estimate. The most common approach is to take the median ORF from all students reaching the comprehension threshold (e.g., 80% comprehension). One problem with this method, however, is that the estimate is sensitive to the sample characteristics, e.g., if data are collected from high-achieving students, benchmark estimates will be higher. Additionally, this approach does not account for the low precision of the estimates that are obtained when only a small sample of students reach the threshold. Therefore, there is a need to develop new methods that produce similar benchmark estimates independently of the ability of students in the sample, and which provide measures of precision for those estimates.

Once such approach has been developed by Room to Read. The approach involves classifying students into two groups: those with at least 80% comprehension and those with less than 80% comprehension. Statistical models (i.e., logistic regression) of the relationship between fluency and the binary comprehension classification are produced, which estimate the level of fluency at which most (>50%) children are reading with 80% comprehension. This approach has a number of advantages, as described below, compared to other methods, such as using the median fluency score of students with 80% comprehension or

[5] See https://dibels.uoregon.edu/docs/techreports/dibels-6th-goals-diagnostic-review.pdf
[6] DIBELS 6th edition benchmarks: https://dibels.uoregon.edu/docs/marketplace/dibels/DIBELS-6Ed-Goals.pdf
[7] Note that DIBELS benchmarks are based purely on levels of reading fluency and are not defined in relationship to student performance on an EGRA-like measure of comprehension. In fact, DIBELS does not include a measure of comprehension until grade 3.

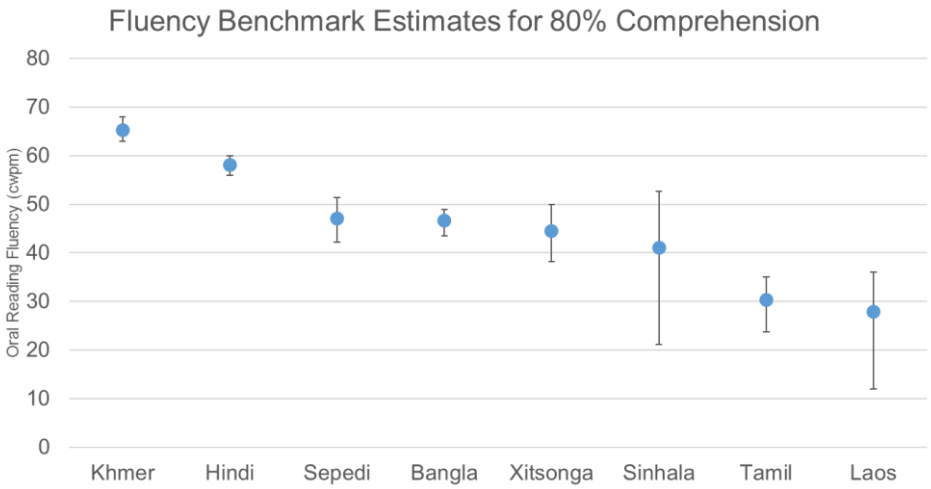
above. If it relies on the EGRA comprehension subtask, it also then inherits the limitations of that way of assessing comprehension.

- Estimates are relatively independent of the ability of students sampled.
- It allows more flexibility in the definition of the benchmark. For example, the method can be used to determine the rate of fluency at which students just begin to understand the passage or can be set at a level where comprehension is universal.
- The method provides an estimate for the strength of the relationship between fluency and comprehension, i.e., whether fluency is a good proxy for comprehension in a particular country's data set.
- Statistical models produce levels of precision associated with the estimate, which serve as a measure of whether the benchmarks are reliable.

Room to Read has piloted this method and is currently producing a more detailed report, systematically comparing this method with various other approaches to defining benchmarks. The report, due to be released in October 2017, will also help inform the best approaches to collecting data to produce reliable benchmarks.

Figure 3 shows Room to Read's estimates using logistic regression models. The graph shows the estimated benchmarks with 95% confidence intervals for eight languages.

Figure 3: Estimated fluency benchmarks (and precision levels) derived from logistic regression analysis



It is apparent from ***Figure 3*** that there is variability in the levels of benchmarks across languages and in the level of precision (indicated by the length of the lines showing the confidence intervals around the point estimates). ***Table 7*** summarizes data that may help explain the variability in levels of precision. The two most precise estimates (smallest confidence intervals) are for Hindi and Khmer. For these two languages, the model fit was good (indicated by the green rather than red color), meaning that there was a strong relationship between fluency and comprehension; the comprehension measures had good internal reliability; the comprehension questions were not affected by ceiling or floor effects (the other four languages had mean scores of between 3.5 and 4 out of 5 questions, suggesting the possibility of ceiling effects); and there was a large sample size.

Table 7: Characteristics determining the precision of fluency benchmark estimates

Language	Model Fit	Comprehension Reliability	Mean score (out of 5)	N
Hindi	0.68	0.94	1.37	1,784
Khmer	0.63	0.80	0.94	1,657
Lao	0.35	—	4.7	180
Tamil	0.26	0.77	3.88	297
Xitsonga	0.12	0.31	3.62	540
Sepedi	0.10	0.41	3.5	540
Sinhala	0.04	0.58	3.51	300

These analyses point to preliminary conclusions about the model fit, reliability and distribution of comprehension scores and sample size that are required for setting precise fluency benchmarks. These conclusions will be tested more systematically in the ongoing Room to Read analysis.

4.9 Data Requirements for Effective Benchmarks

The above analysis and experience of setting benchmarks in a variety of settings leads to numerous guidelines when collecting data for a benchmarking exercise.

1. Sample sizes of less than 200 have typically produced unreliable benchmarks. In general, the larger the sample the more accurate the benchmark.
2. The sample should contain enough students around the level of the benchmark. Samples where very few students reach the comprehension threshold (e.g. 80%) or where almost all students reach the threshold are not useful for benchmarking.
3. Reliable and valid comprehension measures should be used. One rationale for using fluency as a measure of progress is that it can be measured more reliably than comprehension. However, it is essential that the comprehension measure used in the benchmark setting analysis is as reliable as possible. When based on reading a passage, students should be given enough time to read to the end of the passage. Consider assessing comprehension independently from fluency—for example, a separate comprehension subtest based on reading of a second passage. Assess the quality of comprehension measures by assessing the mean and standard deviation of scores on individual comprehension items, as well as the internal reliability of the measure (i.e., how scores on comprehension questions correlate). The comprehension measure should be piloted and assessed before conducting the benchmarking exercise.
4. Use of at least two passages can improve the reliability of the benchmark. Benchmarks can vary with the difficulty level of a passage—both in terms of readability and comprehension. It may be particularly useful to select two passages that are equated for difficulty and are grade appropriate.

Summary: Implications for Asia

Most countries in Asia show higher levels of reading performance in early grades, especially when compared with Africa. One implication of this is that fluency benchmarks are likely to be more useful measures of progress in Asia countries. Because EGRAs that have been administered in Asian countries show a broader distribution of scores (compared to distributions in continents like Africa, which tend to be heavily skewed to low scores or zero), benchmarks in Asia can be set closer to the center of the distribution, with appreciable percentages of students scoring above the benchmark level. In contrast, benchmarks in low performing countries end up being set to the far right of the distribution (with only a small number of students achieving that level of reading proficiency). Benchmarks that are closer to the center of a distribution, like those in Asia, are likely to be more sensitive to improvement and, therefore, more useful for tracking progress. Under such circumstances, a simpler more efficient approach to benchmarking can reliably be taken, e.g., one benchmark for oral reading fluency could serve as an effective indicator. Countries in which students in early grades score at the upper end of achievement distribution may also consider the alternative approach of setting fluency benchmarks based on percentiles of the achievement distribution in a normative sample of students on a national, standardized test, rather than with reference to levels of comprehension.

As discussed in Section 2, the validity of ORF as a proxy for comprehension in alphasyllabaries and logographic languages is based on theory but is not yet well substantiated with empirical findings. Ideally, we recommend using a more robust comprehension measure than what is typical in a standard EGRA. Barring that, we recommend using analyses of the relationship between fluency and comprehension in these language (as presented in ***Figure 1*** for the Philippines or in the above discussion about new approaches to benchmarking) as an opportunity to verify this assumption in each language.

5 Experience of Setting and Using Fluency Benchmarks

In this section, we review RTI and other organizations' experiences in supporting governments to set early grade reading benchmarks. In theory, the process of setting reading benchmarks is a simple matter—it requires selecting a benchmark indicator, reviewing relevant data, gathering key stakeholders, and holding a moderated discussion. In practice, however, each of these steps requires careful planning and a range of important considerations. These considerations are especially important for ensuring that benchmarks are not only set appropriately but also for providing the strongest opportunity for them to be approved and adopted by the government.

Although the themes and best practices noted throughout this report are drawn from a wide range of benchmarking experience across the globe, the remainder of this section relies on specific examples from recent work in the countries shown in ***Table 8*** below.

Table 8: Countries with experience setting benchmarks using the methods described in this paper

Country	Source & Year	Benchmarking Activity	Institutionalization Status
Ethiopia	**EdData in 2010**	**Independent**	**None**
Ethiopia	Reading for Ethiopia's Achievement Developed Technical Assistance in 2015	Intervention-based	Explicitly noted for review/revision in 2017 USAID Ethiopia RFP
Kenya	Hewlett Foundation EGRA in 2010	Independent	None
Kenya	Primary Math and Reading Initiative in 2013	Intervention-based	Adopted for use by Tusome; officially adopted by MOE
Liberia	Liberia Teacher Training Program 2 in 2014	Intervention-based	Currently under review/revision with MOE under Read Liberia program
Malawi	Teacher Professional Development Program in 2014; MERIT: The Malawi Early Grade Reading Improvement Activity in 2016	Intervention-based	Part of government's National Reading Program
Pakistan	Pakistan Reading Project & Sindh Reading Program in 2015	Intervention-based	Officially adopted by MOE
Philippines	EdData and Basa in 2014	Independent	???
Tonga	Tonga and Vanuatu Reading Assessment in 2009	Intervention-based (PEARL)	MOE interest in review but not official adoption
Uganda	Hewlett Foundation EGRA in 2009	Independent	None
Vanuatu	VANEGRA in 2010	Intervention-based (PEARL)	Continued use by development partners

The benchmarking activities in these countries represent work undertaken by Education Development Center, International Rescue Committee, RTI, and the World Bank. All of these experiences have shown that practical benchmarking work follows a consistent 5-step pattern.

1. The aims and scope of the benchmarking activity must be clearly defined.
2. Relevant data must be obtained to address the pre-defined aims.
3. Benchmarks should be set in a participatory workshop that involves representation from a range of stakeholder groups.
4. Short- and long-term targets (based on the newly defined benchmarks) should be agreed upon.
5. Benchmarks and targets should be disseminated to obtain wide-ranging approval and institutionalization.

Accordingly, the remainder of this section is divided into five subsections: (1) Aims, (2) Data, (3) Benchmark setting, (4) Target setting, and (5) Institutionalization.

5.1 Aims

Although the main benchmarking and target setting activities typically occur in a participatory workshop format, it is necessary to discuss clear aims prior to undertaking the standard-setting work. Therefore, the purpose/aim/scope of benchmarks should be discussed with government officials and interested development partners prior to starting of all benchmarking workshops with a broader audience, but should be agreed upon in the beginning of the workshop to ensure participation and buy-in.

5.1.1 Benchmarking the right skills

The first decision to be made at the start of any benchmarking exercise is what benchmarks actually need to be set. In other words, for what skills/constructs are benchmarks required? Section 2 provides the scientific justification for focusing on fluency as a proxy for reading comprehension as an early grade reading benchmark, but how does this align with what countries see as the definition of "reading" and how does that definition drive the benchmarking process?

Since reading is a complex construct, setting a reading benchmark requires careful consideration and a common understanding of what is meant by the term. Can "reading" be defined by a single skill/subtask or does it require a range of both lower- and high-order skills? Are benchmarks required for all components of reading or should certain skills/subtasks take precedence? Should benchmarks be defined by individual skills or composite scores? Experience across a large number of countries suggests that participants have consistently been interested in a benchmark for comprehension and that they feel comfortable reporting fluency benchmarks as their main proxy when using the EGRA. However, some countries have additionally been interested in setting benchmarks for lower-order skills to provide a range of benchmarks that can be used to show learning progression. In Pakistan, for example, benchmarks were set for seven different constructs—ranging from the most basic alphabetic awareness skills (e.g., syllable sounds) to ORF as a proxy for comprehension. In Tajikistan, benchmarks were set in each grade for phonological awareness, dictation, fluency, and comprehension in both Tajik and Russian. In contrast, in the Philippines, benchmarks were only set for comprehension and fluency.

5.1.2 Agreeing on target grades

The majority of EGRA-based benchmarking activities claim to provide benchmarks for early grades. However, the term "early grade" requires specification. For example, in some countries "early grade" may end in grade 4, while in others it may go up to grade 6. Furthermore, "early grade" may refer to just a single grade level or a range of grade levels. When the term refers to a range, participants must decide if benchmarks are required for every grade and whether they can be consistent across grades or if they need to be independently defined for each grade.

To be able to answer these questions, it is first necessary to determine the purpose of the benchmarks themselves (e.g., Who will be the primary users? What will the benchmarks be used for?). If the purpose of setting benchmarks is to provide measures of program performance or reference points against which national assessments can be compared, benchmarks would only be required for those grades and skill areas targeted by such programs and/or assessments. On the other hand, if benchmarks are intended to provide formative measures for school leadership or teachers to judge the performance of their

students as they progress through the learning cycle, it would be necessary to determine measures for all grades, in all skill areas.[8]

In Pakistan, key stakeholders decided that the primary end users would be teachers/head teachers and curriculum developers. Therefore, they determined that benchmarks would be set for all skill areas and individually for each grade from 1 to 5. Conversely, in Tanzania, the Ministry of Education determined that the main purpose of benchmarking was to provide a national diagnostic of system performance. Accordingly, benchmarks were only set for target grades (2 and 4) and focused on the top two higher-order reading skills: ORF and reading comprehension. This was also the case in the Philippines, where EGRA is being used to monitor system level progress and not as a comprehensive assessment of student skill development. The Philippines set benchmarks for reading fluency and comprehension for grades 1–3.

In Kenya, the Tusome project has used the same benchmark, based on the same reading passage, for grades 1 and 2. This has simplified the process and made comparisons more straightforward. Tusome has now created a different benchmark for grade 3, which will shortly be introduced to the program. Jordan has applied a similar approach, using a single benchmark for student performance in grades 2 and 3, with, like in Kenya, the expectation that greater percentages of students in higher grade would achieve the benchmark.

5.1.3 Setting benchmarks for different languages

As noted in Section 2, the relationship between fluency and comprehension differs by language. Accordingly, fluency benchmarks are, by design, language-specific (as the oral fluency rate associated with full reading comprehension cannot be assumed to be the same across languages). Therefore, in any multilingual context, it is essential to determine which languages will have benchmarks set for them and to ensure that appropriate representatives are involved in the benchmark setting activity for all relevant language groups.

The number of languages chosen may result in part from the institutional capacity and availability of teachers, materials, and support for each language. For example, one important factor in the adoption of 2013 Kenya benchmarks was that they were conducted only in the national languages of English and Swahili. The 2010 Kenyan benchmarking exercise also involved the regional languages of Luo and Kikuyu. However, there was less buy-in for these regional language benchmarks because there are no curricular materials in these languages and stakeholders perceived them to have less relevance to the classroom.

Benchmarking exercises in Uganda and Ethiopia were also conducted in local/regional languages. In such cases, it is important to remind participants that benchmarks should not be compared across languages. This may lead to benchmark workshops being held at different geographic levels. In Ethiopia, benchmarking in regional languages required participation of education sector leadership and technical staff from each of the concerned regions. It was evident that each region (or zone for the Southern Nations, Nationalities and Peoples' Region) needed to consider the linguistic characteristics of its language of instruction, the level of development/history of that language as an academic language, and status of their regional education systems. Each region only "bought into" benchmarks that it felt were relevant for and reflective of these contextual factors. By contrast, Pakistan chose to adopt national benchmarks for Urdu and Sindhi, even though the Sindhi data came only from the Sindh province and not from other provinces where Sindhi is spoken. Finally, the benchmarking work in the Philippines focused on four regional languages, with participants from those regions working to define specific levels of fluency and comprehension that the

[8] Note that this kind of school level use of EGRA requires a fundamentally different approach from the focus of this paper to both the assessment and the way its data are processed and used. However, schools and teachers could still validly apply benchmarks established using national data to evaluate how their students are performing.

data indicated were relevant for each language. ***Table 5*** and ***Table 6*** (see Section 3) present the benchmarks set in regional languages in the Philippines and Ethiopia.

5.2 Experience Using Data in Benchmarking Processes

Data are key to setting valid benchmarks. It is important to keep in mind, however, that benchmarking workshops will likely include participants with a range of technical backgrounds and statistical expertise. This is where workshop facilitators need to strike a balance between the "science" and the "art" of setting benchmarks. Data, graphs, and statistical relationships across skills of interest should all be used to provide examples of the science underpinning the benchmarking process. Since benchmarks are language- and typically grade-specific, these data will provide an essential starting point. However, it is important to recognize the expertise of the participants in the room (e.g., language experts, curriculum specialists, and policymakers) and to use their extensive knowledge to settle on benchmarks that are not only scientifically (i.e., supported by data) defensible but also politically feasible, aligned with learning expectations, and ultimately "make sense". Helping broker dialogue that considers these factors, as well as what the data say, is what constitutes the art of benchmark setting. Additionally, prior work has shown that providing a variety of options (i.e., data-based methods) and allowing for flexibility in choosing the preferred approach ultimately leads to greater participation and a sense of ownership over the process and the benchmarks themselves.

Although data are extremely helpful in starting the process, the absence of data proving strong relationships across skills does not automatically mean that no benchmarks can be set. In Pakistan, it was determined that benchmarks should be set for alphabetic awareness. Despite the fact that the data showed weak relationships between alphabetic awareness and higher-order skills (such as ORF and reading comprehension), curriculum and pedagogical experts determined timelines and criteria for mastery of both letter names and letter sounds based on their extensive knowledge of the context.

In many countries, data may exist for only certain grades (e.g., 2 and 4), but benchmarks are still expected to be set for all primary grades (e.g., 1 to 5). In these cases, the process revolves around first setting benchmarks in grades where data are available, followed by extrapolating to intervening grades.

For example, at the benchmarking workshop in Ethiopia in 2015, the available data came from an EGRA administered in grades 2 and 3 in 2014. Participants first worked on benchmarks for grade 3 and then for grade 2. For grades 1 and 4, for which no EGRA data were available, the groups had to extrapolate from the benchmarks they had just proposed. To do so, they calculated the difference in performance on the 2014 EGRA for students in grade 2 and those in grade 3. This "grade-to-grade growth" gave them a basis for discussing how much students could improve from grade 1 to grade 2 and, therefore, how the benchmark for grade 1 should be relevant to grade 2. They used the same process to extrapolate from grade 3 to grade 4. An interesting discussion ensued regarding whether one should expect growth in reading ability in early grades to be linear (e.g., with higher growth occurring early, perhaps between grades 1 and 2, and then leveling off as students reach high levels of fluency when moving from grades 3 to 4). The team chose to take a non-linear approach.[9]

Additionally, availability of reading performance data by subgroups (in combination with the pre-defined purpose) can be used to determine the level of disaggregation of the benchmarks. However, it is suggested that regional/local benchmarks are set at the same level as national benchmarks (assuming they are for the same language) but that targets are adjusted based on available data for baseline performance. This is the approach Pakistan

[9] Note that data from benchmarks used in the US—from DIBELS—support the assumption of non-linear grade-to-grade changes in reading performance.

took to simplify the process, with an understanding that the relationship between skills does not change by region but that the percentage of children at given performance levels does.

Lastly, it is important to keep the limits of data in mind throughout the process. For example, in some countries there may only be a small proportion of students reading with comprehension. Therefore, if one were to rely on these data alone, benchmarks would be set based on the performance of a small number of observations. For example, while there were more than 6,600 students who took the 2013 national EGRA in a local language in Ghana, only 46 of those students scored 80% on the comprehension subtask (and half of them were from a single region). Therefore, the precision of the related fluency score would be very low, as it would be based on less than 1% of the total data. In these situations, it may be important to test alternative model specifications to account or adjust for the small sample sizes (as discussed in Section 3.8). As noted in Section 3 in higher performing countries, data tend to provide more reliable estimates of relationships between and across skills and, therefore, those data can be given more weight in the process.

5.3 Participatory Approach to Benchmark Setting

Once the aims of the benchmarking process are defined and the appropriate data are available, a benchmarking workshop should be organized to set the actual benchmarks. This activity is broadly made up of two main components: (1) gathering key stakeholders and (2) holding a moderated discussion.

5.3.1 Gathering key stakeholders

Perhaps the single most important aspect of securing buy-in and adoption of benchmarks is ensuring that the right people are working together during the benchmark setting process. This entails inviting participants from a range of departments and institutions, with a range of responsibilities and oversight. At the very least, a benchmarking workshop should include (as appropriate) decision-makers from the central ministry of education, the department of curriculum, the department of teacher training, the department of school supervision/inspection, the department of education policy/planning, the examinations/assessment board, national and regional directors, language experts (within the ministry and/or academic institutions), as well as representation from teachers and school leaders.

High-level officials' participation should not be limited to the benchmarking workshop itself. As previously noted, it is necessary to have clear discussions about the purpose and use of benchmarks with high-level government officials prior to the benchmarking workshop and to allow for continued conversations and dissemination of benchmarks after the workshop. Only with a strong sense of ownership, understanding, and acceptance will benchmarks have an opportunity to be institutionalized.

The importance of gathering the appropriate stakeholders is demonstrated in an example from Kenya. Key officials to be involved in benchmarking work in Kenya included (at a minimum), representatives from the assessments/examinations unit, the director of policy, the national body responsible for teacher employment and training, and regional leaders. The 2013 Kenya benchmarking exercise involved such participants but the 2010 exercise did not. Additionally, the 2013 benchmarking exercise in Kenya was set up for success both by holding a brief Ministry technical staff meeting before the benchmarking workshop and by involving Ministry technical staff in data collection. As a result of these high-level collaborations and buy-in, the 2013 benchmarks were officially adopted by the Kenyan MOE, while the same was not possible for the ones developed in 2010.

The added challenge of having the right decision-makers in the same workshop in a country where benchmarks are being developed for regional languages is evident in the experience of the Philippines. At a benchmarking workshop in 2014, the Philippine Department of Education (DepEd) assigned staff from the office of the Undersecretary for Programs and

Projects, the Bureau of Elementary Education, and from six regions to attend the workshop. The regions that participated included Region I, Region IV-A, the National Capital Region, Region VI, Region VII, and the Autonomous Region of Muslim Mindanao (ARMM). A total of 49 participants spent a day examining data from the EGRA surveys conducted in 2013 and 2014 and used those data to propose benchmarks for reading performance in grades 1–3 for Filipino and English and for four regional mother tongues: Ilokano, Hiligaynon, Sinubuanong Binisaya (Cebuano), and Maguindanaoan.

The group recognized that the DepEd central office would need to be the final arbiter of the standards—validating and strengthening, where necessary, the proposed benchmarks. However, the group of participants also recognized the need to enlist the support and input of regional management committees, as well as other regional-level technical staff. For ARMM, given its particularly autonomous status, the group was concerned that the Regional Secretary would need also to review, comment on, and officially accept any standards. Ultimately, the complexity of obtaining first agreement and then official adoption of the language-specific benchmarks across all the interested authorities and stakeholders kept these benchmarks from being officially adopted by the country. Although the DepEd decided to continue to use EGRA to monitor improvement in early literacy acquisition at the system level (with regional disaggregation), it has yet to adopt benchmarks for reading performance based on EGRA measures of ORF (or for any other skill area).

5.3.2 Holding a moderated discussion

In addition to helping clearly define a purpose, analyzing the data, and gathering key stakeholders for a workshop, the organization in charge of leading the benchmarking process (from a technical standpoint) should also be prepared to moderate the discussion around benchmark setting. This requires (or is greatly strengthened by) prior experience in benchmarking, expertise in early grade reading data (and analysis), understanding of education policy/systems, and knowledge of the roles and relationships among the participants. If this process is being led by an external organization, it is essential that they can provide technical expertise and assistance without steering the process toward a pre-determined outcome. The participation and ownership from the participants must also be carefully balanced with the need to arrive at consensus by the end of the workshop, which often requires a lot of careful negotiation and compromise.

A standard approach to conducting this moderated discussion is illustrated by the process in Vanuatu and Tonga. In both countries, discussions with the Ministry of Education began with presentation of data and focused on the scatter plot of fluency against comprehension. These data facilitated a conversation among policy makers about an acceptable level of comprehension. Once that conversation was concluded, the scatter plot was used to determine the fluency level indicative of acceptable comprehension levels.

Other countries have had different experiences with benchmarking. In Liberia in 2014, the benchmarking workshop assembled a large group of stakeholders from the Ministry of Education, from academia, and from many of the government's implementing partners. The broad participation enabled the benchmarks to be widely shared and recognized; however, it also meant that the involved stakeholders had to agree to employ a process of arbitration when groups had divergent opinions about what was a reasonable benchmark to set. Interestingly, while there were some differences in the level of skill acquisition at which different actors believed the benchmarks should be set, most of the debate was around what targets were reasonable to set – owing primarily to wide variation in the degree to which different stakeholders had any confidence that the education system could work to improve instruction sufficiently enough to help large numbers of students meet benchmark performance.

The Kenya meeting in 2013 was successful in part because genuine decisions were being made based on several viable benchmarking options, presented by the lead organization. It

was possible to get key policy makers together for the meeting only by keeping it short, which in turn required sound preparation. Technical staff were involved in data collection, which helped save time establishing the validity of the data during the benchmarking meeting. Also, analyses were presented to technical staff in advance to avoid lengthy explanations during the benchmarking meeting. Several analysis methods were used to produce different benchmarks, while fluency levels associated with non-zero comprehension and 75% comprehension were presented. One aim of the meeting was to set benchmarks from these multiple options because in the previous exercise in Kenya, facilitators had favored one method of benchmarking, which led to less buy-in than when the decision-making was left up to the participants.

5.4 Target setting

Although benchmarks represent standards of learning achievement against which student performance can be judged, targets are designed to provide estimates for the percentage of students expected to meet benchmarks by a given time point (e.g., 1 year, 5 years, or 10 years). Accordingly, targets provide a means of evaluating progress in reading performance (either by a program or an education system). The process of setting targets is similarly a combination of science and art.

After benchmarks have been set, existing data should be used to determine the current (or baseline) percentage of students meeting each benchmark. This is the "science" and it provides a starting point for the discussion on targets. In an ideal scenario, data would be available showing the impact of a reading improvement intervention on the skill(s) and in the grade(s) in question, which would provide an estimate of the type of reading improvement that could be expected if the intervention were continued, adapted, or brought to scale. In both Liberia and Malawi, where data were available from interventions in each country, participants in the benchmarking discussions looked at the amount of improvement the data showed was possible in each case. In Malawi, because data were available from two national EGRAs (in 2010 and 2012), participants considered how scores improved during that period, as shown in ***Table 9***.

Table 9: Malawi's national EGRA results in 2010 and 2012

Skill Area	Standard 2	
	2010	2012
Letter name knowledge (correct letters per minute)	**2.3**	**5.7***
Syllable reading (correct syllables per minute)	**1.4**	**3.2***
Familiar word reading (correct words per minute in isolation)	**0.8**	**1.9***
Non-word reading (correct non-words per minute in isolation)	**0.5**	**1.2***
ORF (cwpm of text)	**0.8**	**1.3***
Oral reading comprehension (# correct out of 5 questions)	**0.0**	**0.0**

**indicates a statistically significant difference (p<0.05) in the means for 2010 & 2012.*

Given the modest changes in skill levels between 2010 and 2012 at the national level, these data made participants somewhat pessimistic about the targets that they could realistically set for students to meet the benchmarks for proficient skill acquisition.

However, additional data were also available from a pilot reading improvement effort in two districts.[10] Those data showed considerably more improvement was

[10] The pilot program referred to was implemented in two districts by USAID's Malawi Teacher Professional Development System program.

possible when teachers and students were the beneficiaries of an intervention. For example, in the pilot program, ORF increased from an average of 0.5 cwpm to more than 7 cwpm (compared to the national increase from 0.8 to 1.3). Given those data, the participants argued for higher targets than they may have otherwise settled on. Ideally, distributions of performance (as well as changes in zero scores) should be reviewed alongside mean results. However, mean results do at least provide an illustrative example of the magnitude of overall performance gains that align with most donor-required indicators.

The second-best option would be to have multiple years of performance data, in which case past trends in educational performance could be calculated and used to provide a basis for determining future growth in student performance. In the absence of data from multiple time points, single-year data could be used to estimate differences in performance across grades, which could then serve as the basis for understanding how much improvement to expect from one grade (or one year) to the next. Regardless of what type of data are available, the most complicated part of target setting is for all participants to agree on the expected improvements in performance over time.

Although the data provide important starting points, there are a few considerations that require nuanced discussion and compromise. These conversations constitute the “art”.

5.4.1 Benchmarks inform targets, not the other way around

By the time participants have reached the discussion on targets, benchmarks should already be set. However, in settings where baseline data show that few (or a small percentage) of students are currently meeting benchmarks, there is often a desire to revisit the benchmarks themselves. The argument is that since so few students are meeting the benchmarks, they must have been set too high. It is important to remind participants that the benchmarks were set based on a combination of relationships in the data and participant expertise on the education system and learning expectations. Therefore, benchmarks should represent the actual desired level of skill acquisition, and targets should be realistic based on assumptions about how much improvement is achievable. Lowering the benchmark to increase the current number of students meeting that benchmark undermines the value of the benchmark itself. For example, it is significantly more valuable to know that only 5% of students are reading with a level of fluency that will allow them to read with comprehension, than it is to know that 50% of students are reading with an arbitrarily lower level of fluency that no longer corresponds to comprehension or higher-order skills.

An example of this problem was seen in Kenya, where targets were initially set for each benchmark. Only 6% of children were reaching one of the benchmarks, which had a target of 40%. Project staff stated that it was uncomfortable to push and talk about the 6%, so the targets were not used. Concerned stakeholders in Kenya only tracked the percentage of children reaching the benchmark without reference to the target. This kept the meaningful benchmark intact but undermined the purpose of the target.

Similarly, the government of Tonga decided to de-emphasize benchmarks and targets in its reading improvement strategy. When it became clear that only 20% of readers were passing fluency benchmarks it was decided that their targets would likely not be achieved in the immediate future and that the discussion of targets may distract from efforts to improve classroom instruction. As one staff member observed, they got caught up in the setting of benchmarks[11] before focusing, instead, on strategies needed to improve reading and understand what is achievable in the short term, and at each stage of education reform.

In the Philippines, on the other hand, benchmarking workshop participants were comfortable setting targets, without considering compromising the benchmarks they had set based on the given relationships between fluency and comprehension in each language. They expressed confidence when setting ambitious targets, primarily because the DepEd was in

[11] Personal communication to one of the authors.

the midst of implementing and supporting its curriculum reform emphasizing maternal language instruction in the early grades.

One alternative to ensure that benchmarks/targets represent genuine learning expectations, while also being politically viable, would be to set various levels of benchmarks instead of using a single cut-point. For example, in the Philippines, benchmarks were defined not only for proficient readers (i.e., those meeting agreed-upon standards for skill mastery) but also for emerging readers (i.e., those who could read below the level of mastery but above the non-reader category).

In Pakistan, ORF benchmarks were divided into three categories per grade: (1) does not meet grade level expectations, (2) meets grade level expectations, and (3) exceeds grade level expectations. Meeting expectations was defined by examining the range of ORF scores that is associated with full comprehension. For example, the grade 4 measure for meeting expectations was 80 to 120 cwpm in Urdu. This range was closely aligned with the 25th to the 75th percentile of ORF scores among students scoring 80% or higher on reading comprehension. Ultimately, performance standards were all established by examining the spread of scores from the 1st to 3rd quartiles (i.e. 25th to 75th percentiles) and selecting an appropriate and acceptable range that represented the results of "average" students in that range. Not meeting expectations was defined as being below the lower threshold of average performance (i.e., below 80 cwpm in grade 4 in Urdu) and exceeding expectations was defined by those students at the very top of the range (i.e., above 120 cwpm in grade 4 in Urdu). This allowed for targets to be set at three different levels for each grade, which lessened the impact of having a smaller proportion of learners at the highest level.

Similarly, in Zambia, using a range of categories (non-readers, emergent readers, and readers) allowed for a more honest conversation about the current (and future) low levels of performance at the reader level. Progress was instead defined more heavily by a focus on reducing non-readers while keeping a more reasonable and modest target for the readers.

Although the majority of sub-Saharan African countries have small percentages of students reaching benchmarks, the proportions of students meeting benchmarks in countries in Asia is expected to, and likely will, be significantly higher. This should reduce concerns about "not enough" students meeting standards at baseline, particularly for higher-performing countries. Benchmarking experience thus far has confirmed that for countries in Asia, general performance is higher and there is more of distribution of students scoring across a fuller range of levels of reading fluency (e.g., not nearly as large a percentage scoring zero as in sub-Saharan Africa). Therefore, the tension around whether enough students are shown to be meeting the benchmark has been less of an issue.

In all cases, what has been apparent from the experience of helping countries set benchmarks is that the process leads education decision makers and other stakeholders to have an honest conversation about student performance – and to do so in terms of tangible measures of skill development, not just in terms of "passing" or "failing" grades or exam scores.

5.4.2 Projects versus systems

Although intervention data are most appropriate for setting future performance targets for a project or program that has similar components to the intervention, these data can also be used to provide an estimate of expected improvement for an entire education system. However, it is important to keep in mind that data from carefully implemented small- and medium-scale interventions are typically considered to be "best case scenario" results. In other words, it is likely unrealistic to assume that an education system, as a whole, will be able to produce the same improvements as a targeted intervention project.[12] This is

[12] For an analysis showing the drop off in effect sizes between pilot and scaled up implementation of interventions see Moore, Gove, & Tietjen, 2017.

especially true when there is not significant buy-on and ownership of the approach by the government. Therefore, system-level targets (as well as targets that extend beyond the scope of an intervention) should be adjusted downward to account for the more difficult nature of making improvements at a larger scale and without the extensive targeted support of an intervention project. There is no specific rule of thumb for exactly how much the targets should be reduced, but this decision should be made in consultation with stakeholders who have the most thorough understanding of system constraints and expectations.

A good example of this is evident in the data used in the benchmarking workshop in Liberia. Data from the initial, small scale pilot program (EGRA+) were compared with data from a program that implemented a more than six-fold expansion of that model (LTTP2). ***Table 10*** shows the difference in the gains that were realized in three skill areas by EGRA+ compared to LTTP2, showing much lower outcomes for non-word reading, ORF, and comprehension when the program was expanded from 120 schools to 792 schools. The discussion that ensued about what one could assume would be the case if a similar intervention was to be taken to full scale in Liberia was lively, with honest appraisal of what could be expected given the country's existing institutional capacities in the education sector.

Table 10: Comparison of pilot and expanded implementation results in Liberia

	Grade 2	
EGRA+	**Baseline**	**Midterm**
Non-word reading (cwpm in isolation)	**1.4**	**13.1**
ORF (cwpm of text)	**15.0**	**43.2**
Oral reading comprehension (# correct out of 5 questions)	**0.9**	**2.4**
LTTP2		
Non-word reading (cwpm in isolation)	**0.3**	**3.2**
ORF (cwpm of text)	**4.8**	**14.2**
Oral reading comprehension (# correct out of 5 questions)	**0.3**	**0.7**

This example demonstrates that the determination of benchmarks and setting of targets for the percentages of children that will meet them in the future cannot and should not be divorced from discussion about and commitment to what efforts are underway or will need to be undertaken to improve instruction. Targets are not going to be met simply because of the passage of time—in all systems, well-implemented interventions are needed if significant improvements in outcomes are going to be realized.

5.4.3 Feasibility versus high expectations

Feasibility and high expectations are not diametrically opposed to one another in theory, but they are often at odds in practice. A balance must be struck between setting targets that will push the system to improve and those that can be attained. This plays out most commonly in project-based target-setting exercises. Funding agencies seek targets that are high enough to justify the cost of the project, ministries of education want targets that will show impressive gains in performance (particularly when baseline levels are low), and project implementers need targets that will be attainable so that they do not set themselves up for failure from the

start. Accordingly, coming to agreement on targets is no easy task. Often, initial suggestions for targets in participatory workshops will be far higher than can reasonably be attained. In our experience setting benchmarks and targets, it has not been uncommon for baseline estimates to show less than 10% of students meeting ORF or comprehension benchmarks and for workshop participants to propose five-year targets upwards of 60%–70% (e.g., Jordan, Nepal, Pakistan, Philippines, and Tanzania). Although the impetus for these target suggestions comes from a good place (i.e., wanting to show strong improvements in the education system), it is essential to remind participants that targets should be set based on realistic expectations of what can be achieved, as opposed to unfounded desires for where the system may one day be.

Our experience has also shown us that education stakeholders' degree of confidence in the education ministry's ability to successfully implement proposed or ongoing reforms exerts a strong influence on how ambitious or cautious said stakeholders are when setting targets. For example, in the Philippines, stakeholders were confident in the DepEd reforms and in the ability of the system to carry them to fruition. Participants were similarly confident in Pakistan (based on the continued Pakistan Reading Project's [PRP's] work and the ability of the ministry to continue the work after program completion). Conversely, Liberia and Malawi had different approaches. In Malawi, questions concerning institutional capacity exerted downward pressure on the targets that participants ended up agreeing on. Encouragingly, they did not compromise on the benchmark for proficient reading, but they did set low expectations for how many children would be able to meet it in the near term. In fact, that outcome represents the ideal scenario: Benchmarks that are realistic in that they represent a level of reading skill commensurate with children being able to read and understand grade-level text but targets that are also realistic and do not expect dramatic change in a relatively short period of time given the institutional challenges that most education systems face.

It is also important to remember that while targets provide estimates of expected learning improvements, it is possible and advisable to revisit and revise targets based on the availability of new data and/or changes to the educational landscape (whether for better or for worse).

5.5 Institutionalization

As noted above, the setting of benchmarks and targets requires several key steps and considerations but, if followed, the process itself can be relatively straightforward. More complicated, on the other hand, is the ultimate adoption and continued use of the benchmarks once they have been set. The first step toward adoption of benchmarks is involving key stakeholders throughout the process (from inception meetings through the benchmarking workshop to dissemination). It is valuable to begin the benchmarking conversation as early as possible (even well before the benchmarking process in longer-term projects), in order to ensure that the government has a full understanding of how benchmarks will be set, how they can be used, and what value they have for improving education systems. Additionally, it is important to understand the continued role that donors, ministries, and projects all must serve once the benchmarks have been set. While governments are ultimately responsible for ensuring that benchmarks are institutionalized, donors and projects are often required to play a continued role in providing financial and technical support, as well as providing clear guidance on how benchmarks can be used to continually improve education. In fact, a project can play a critical role in modeling how benchmarks can be used to evaluate system improvement over time, thus reinforcing for ministry counterparts the value of institutionalizing benchmarks. Additionally, as subsequent rounds of data become available, a project can support the ministry in revisiting and re-evaluating benchmarks. For example, introducing and using a better measure of comprehension than the typical EGRA comprehension subtest could help a ministry develop more reliable benchmarks.

A project promoting the use of fluency benchmarks increases the likelihood of them being officially adopted. Under the Kenya Tusome program, fluency benchmarks were widely disseminated and discussed (including the presentation of benchmarks at meetings with teachers and coaches throughout the program). In Vanuatu, the use of learning standards is continued, at the least, by development partners who include it in their reports. In Tonga, there is continued support for taking "the pulse of the national education system" with a transparent measure and a reference point. In Pakistan, benchmarks have been adopted but the government has noted that they will need continued project support in tying them to national standards/curricula for true institutionalization. Conversely, a lack of project support helps explain why benchmarks were not adopted from other benchmarking exercises in similar countries. For example, in Ethiopia in 2010, benchmarks were created for Amharic and Tigrinya reading, but momentum was lost after a delay starting a USAID-funded project, which would have helped apply those benchmarks (thus, why benchmarks were revisited in 2015).

In Pakistan, a clear-cut policy has been drafted to officially accept the benchmarks that were set in 2015. This shows a significant amount of support from the Ministry of Education; however, acceptance does not automatically lead to continued use. For example, although standards were set for grades 1 to 5, the PRP only works in grades 1 and 2, leading to concern about how teachers in grades 3 to 5 will be taught about recognition and implementation of the standards. Ultimately, the claim is that standards will not be practically workable without either project support or integration into official government curriculum, textbooks, and professional development opportunities. In Pakistan, as in many other places, teachers feel responsible for material in their textbooks/curriculum and teach what they are trained to teach. Therefore, working toward benchmarks/standards without integration into the education system seems unsustainable once the project ends.

Another factor in the continued use of benchmarks is clearly understanding the intended audience. Both Tonga and Vanuatu downplayed the use of the term "benchmarks" and instead discussed "reference reading standards to monitor reading development in the early grades" (Vanuatu) and "indicative learning milestones" in Tonga. In each case, the use of the milestone/standard was to help teachers support learning in the classroom and monitor how quickly children progress through letters, words, and passages. It was not used as a high-stakes test. This decision was made in light of experience in Cambodia, which showed that the more benchmarks that were emphasized, the more pressure teachers felt.

Ultimately, a country's adoption and continued use of benchmarks relies on buy-in and ongoing work on the part of all stakeholders (including dissemination from projects, support from donors, and institutionalization from ministries).

Summary: Implications for Asia

By and large, the preceding process for setting benchmarks and targets is recommended for all contexts and levels of performance. Although much of the experience described in this section came from African countries, the process of clearly defining aims, gathering relevant data, using a participatory approach (with key stakeholders) to set appropriate and achievable benchmarks and targets, and disseminating results for approval and institutionalization, are all integral to the success of benchmark setting and continued use in Asia.

There are, however, two aspects that may differ in Asian countries with higher levels of reading performance. First, as noted in Section 3, a larger proportion of students reading with full comprehension will provide more precise estimates of benchmarks. This will serve to remove some of the guesswork from the benchmark setting process (i.e., the stronger relationship between fluency and comprehension will lead to more agreement about where the benchmark should be set). Second, having a greater percentage of students meeting the benchmark at baseline should help minimize the desire to inappropriately lower benchmarks (due to concerns about "not enough" students meeting standards initially). This can lead to a more streamlined and less contentious benchmark and standard setting process.

6 Conclusions and lessons learned

6.1 Conclusions about the Science of Language Development and Assessment

Our review of the science of language development aimed to understand how existing research, predominantly on alphabetic languages, such as English, applies to nonalphabetic languages in Asia, such as in alphasyllabaries (e.g., Korean, modern Lao, and Khmer script) and logographic languages (e.g., Chinese and Japanese kanji). A high-level conclusion is that there are many similarities in the process of learning to read across languages. The same foundational processes are important in alphabetic, alphasyllabic, and logographic languages. Thus, the science of reading supports the assumption that EGRA assessments, and the use of fluency benchmarks, are appropriate to use across Asia.

Reading fluently is a key component of reading proficiency and is also a proxy for reading comprehension. This relationship has been established in alphabetic languages, particularly English. Some data suggest the relationship is maintained in alphasyllabic and logographic languages. However, we recommend that future benchmarking exercises assess the strength of this relationship in the assessment data being used in each exercise.

Language characteristics can also influence the rate at which children learn to read. For example, learning to read in Chinese is a much slower process than in alphabetic languages. There may be a need to set more modest targets in more complex languages and, as with all benchmarking exercises, resist the temptation to make direct comparisons between languages.

Another element to keep in mind is that counting words in logographic or alphasyllabic languages can be challenging because of ambiguities in word boundaries. Some approaches to this issue involve counting characters or syllables rather than words, focusing on errors in word segmentation rather than accurate word segmentation or convening experts to adjudicate on the count of words.

Finally, the recommendation that different benchmarks be set for different languages applies also to different versions of the same language, such as those with and without diacritics. This is particularly relevant to Arabic and some South Asian languages.

6.2 Conclusions on Data Use for Benchmarking

Data collected for benchmarking exercises should involve a sufficient sample size to ensure that an acceptable number of children are above and below the comprehension threshold (e.g., 80% comprehension). Comprehension measures used should have good internal consistency; ideally, two or more comparable passages should be used to reduce variability in benchmark estimates. When possible, fluency measures should be separated from comprehension measures (to ensure the reliability and independence of both estimates). Data reliability is paramount, and data that are unable to meet agreed-upon standards for reliability and validity (such as those in the EGRA toolkit) should not be used for setting benchmarks.

Benchmarks defined in Africa have often proved to be well above the average reading level in each country. Therefore, available data may show very few students meeting those benchmarks (as shown in ***Table 6***) and intervention projects lift few students above the benchmark (***Figure 2***). The disparity between the fluency benchmark and the mean performance of students makes benchmarks less meaningful as a means of tracking student performance. We anticipate that most Asian countries will not have this problem because the mean reading achievement is higher than in Africa. We expect that benchmarks in Asian countries will better enable those countries to reliably track progress in a meaningful way. The highest achieving countries could consider setting normative benchmarks (as is done in the US), based on the distribution of fluency scores, rather than solely with reference to comprehension.

6.3 Conclusions about the Process of Benchmark Setting

The experiences reviewed for this report lead us to conclude that countries are interested in having benchmarks and that most stakeholders readily see benchmarks as a useful means to track system-level performance and progress over time. Nevertheless, demonstrating the reliability of the data used in benchmarking, and providing measures of the validity of the relationship that underpins the benchmarking process (between fluency and comprehension), are critical to ensuring broad acceptance of benchmarks. Making the case for the reliability of the data also increases the likelihood that the proposed benchmarks will be officially adopted and used.

Deciding whether to set benchmarks for every grade and for every skill area assessed depends primarily on their intended use. To track progress at the system level, having one highly reliable indicator of performance, such as the percentage of students meeting the benchmark for ORF, is all that is needed. This is especially true in those Asian countries where performance is higher. In such cases, a reading fluency indicator is likely to be strongly associated with comprehension, sensitive to smaller increments of improvement, and meet the criteria for precision and reliability. If benchmarks are intended to be used by teachers to monitor their students' progress, then defining expected levels of performance for each grade and in each skill area would be more appropriate, especially in the first few years of primary school when foundational literacy skills are developing.

When setting benchmarks and targets, it is necessary to strike the appropriate balance between science and art. The science employed in analyzing the data needs must be evident to the involved stakeholders, with appropriate justifications and explanations (in non-statistical terms that are accessible for all audiences). The art involves guiding dialogue to facilitate the use of data (often by people not accustomed to working directly with data) to animate the dialogue and broker agreement among, sometimes, divergent viewpoints. Our experience indicates that good facilitation depends on trusting the process, i.e., letting

participants spend time digesting the available data, make competing arguments, and work their way to a consensus determination.

We have seen that reaching agreement on the benchmark itself is often less contentious than determining the target for how many children will meet that benchmark in the future. The desired outcome is to have a “true” benchmark that represents a meaningful level of reading ability, as well as a realistic, obtainable target. Of vital importance is to help stakeholders (including development and implementing partners) resist the temptation to lower the benchmark so that a higher target can be more easily reached; this trade off actually subverts the fundamental reason for setting benchmarks. If countries find ORF benchmarks intimidatingly high, one solution is to set benchmarks for lower-level skills, such as letter reading fluency, or set intermediate ORF benchmarks. Such intermediate benchmarks can help track progress that may not be apparent for comprehension-related benchmarks.

A last point concerns the increased importance of setting benchmarks now that the Sustainable Development Goal (SDG) for education includes an indicator (4.1.1) on the proportion of children achieving at least a minimum proficiency level in reading (and math). The global dialogue about SDG indicator 4.1.1 has recognized the linguistic and orthographic differences across languages (as discussed in this paper) and recognizes the different development levels of each country’s education system. Therefore, it is accepted that each country determines its own definition of “minimum proficiency.” The analyses and processes described in this paper for setting benchmarks are intended to help countries do exactly that. With defined benchmarks, countries can then measure and report on their progress in meeting the education-related sustainable development goals.

References

Abadzi, H. (2011). *Reading fluency measurements in EFA FTI partner countries: Outcomes and improvement prospects.* GPE Working Paper Series on Learning, No. 1. Available from http://documents.worldbank.org/curated/en/925221468179361979/Reading-fluency-measurements-in-EFA-FTI-partner-countries-outcomes-and-improvement-prospects

Bartlett, L., Dowd, A. J., & Jonason, C. (2015). Problematizing early grade reading: Should the post-2015 agenda treasure what is measured? *International Journal of Educational Development, 40*, 308–314. doi:10.1016/j.ijedudev.2014.10.002

Basaran, M. (2013). Reading fluency as an indicator of reading comprehension. *Educational Sciences: Theory & Practice, 13*(4), 2287–2290. https://doi.org/10.12738/estp.2013.4.1922

Bastien-Toniazzo, M., & Jullien, S. (2001). Nature and importance of the logographic phase in learning to read. *Reading and Writing, 14*(1–2), 119–143.

Bulat, J., Dubeck, M., Green, P., Harden, K. K., Henny, C. E., Mattos, M. L., Sitabkhan, Y. (2017). *What we have learned in the past decade: RTI's approach to early grade literacy instruction*. (RTI Press Publication No. OP-0039-1702). Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2017.op.0039.1702

Burchfield, S., Hau, H., Baral, D., & Rocha, V. (2002). *A longitudinal study of the effect of integrated and basic education programs on women's participation in social and economic development in Nepal*. Funded by the United States Agency for International Development and the Office of Women in Development. New York: World Education, Inc.

Carnine, D. W., Silbert, J., Kame'eunui, J., & Tarver, S. G. (2014). *Reading development: Chall's model*. Available from https://www.education.com/reference/article/Chall-model-reading-development/

Chall, J. (1983). *Stages of reading development.* New York: McGraw Hill.

Chang, L., Plaut, D. C., & Perfetti, C. A. (2015). Visual complexity in orthographic learning: Modeling learning across writing system variations. *Scientific Studies of Reading, 20*(1), 64–85. https://doi.org/10.1080/10888438.2015.1104688.

Cho, J. & Chen, H. (1999). Orthographic and phonological activation in the semantic processing of Korean Hanja and Hangul. *Language and Cognitive Processes*, *14*(5–6), 481–502. http://dx.doi.org/10.1080/016909699386167

Daane, M., Campbell, J., Grigg, W., Goodman, M., & Oranje, A. (2005). *Fourth-Grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading.*

Dubeck, M. & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. International Journal of Education Development, 40 (2015), pp. 315–322.

Dynamic Measurement Group, Inc. (2010). *DIBELS® Next benchmark goals and composite score.* Available from https://dibels.org/papers/DIBELSNextBenchmarkGoals.pdf

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., van Daal, V. H. P., Polyzoe, N, Tsipa, M., & Petalas, M. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly, 39*(4), 438–468. https://doi.org/10.1598/RRQ.39.4.5

Frith, U. (1985). Beneath the surface of developmental dyslexia. *Surface dyslexia*, *32*, 301–330.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*(3), pp. 239–256. https://doi.org/10.107/S1532799XSSR0503_3

Good, R. H., III, Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading,* 5(3), 257–288. doi:10.1207/S1532799XSSR0503_4

Graham, B. E., & van Ginkel, A. J. (2014). Assessing early grade reading: The value and limits of 'words per minute.' *Language, Culture, and Curriculum, 27*(3), 244–259. https://doi.org/10.1080/07908318.2014.946043

Hogaboam, T. W., & Perfetti, C.A. (1975). Lexical ambiguity and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 14*(3), 265–274. https://doi.org/10.1016/S0022-5371(75)80070-3

Hudson, R., Pullen, P., Lane, H., & Torgesen, J. (2009). The complex nature of reading fluency: A multidimensional view. *Reading and Writing Quarterly,* 25, 4–32.

Jackson, N. E., & Coltheart, M. (2001). *Routes to reading success and failure: Toward an integrated cognitive psychology of atypical reading*. Philadelphia, PA: Psychology Press, 2001.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719–729.

Jukes, M. C. H., Cummiskey, C. P., Gargano, M. N. and Dubeck, M. M. (in press). Data-Driven Methods for Setting Reading Proficiency Benchmarks. San Francisco: Room to Read

Kim, Y. S. G. (2015). Developmental, Component-Based Model of Reading Fluency: An Investigation of Predictors of Word-Reading Fluency, Text-Reading Fluency, and Reading Comprehension. *Reading Research Quarterly,* 50(4), 459–481. doi:10.1002/rrq.107

Kim, Y. S., Park, C. H., & Wagner, R. K. (2014). Is oral/text reading fluency a "bridge" to reading comprehension? *Reading and Writing,* 27(1), 79–99. doi:10.1007/s11145-013-9434-7

Kim, Y. S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does Growth Rate in Oral Reading Fluency Matter in Predicting Reading Comprehension Achievement? *Journal of Educational Psychology,* 102(3), 652–667. doi:10.1037/a0019643

Kim, Y. S. G., & Wagner, R. K. (2015). Text (Oral) Reading Fluency as a Construct in Reading Development: An Investigation of Its Mediating Role for Children From Grades 1 to 4. *Scientific Studies of Reading,* 19(3), 224–242. doi:10.1080/10888438.2015.1007375

Kim, Y-S.G., Wagner, R. K., & Foster, E. (2011). Relations among oral reading fluency, silent reading fluency, and reading comprehension: A latent variable study of first-grade readers. *Scientific Studies of Reading: The Official Journal of the Society for the Scientific Study of Reading, 15*(4), 338–362. https://doi.org/10.1080/10888438.2010.493964

Kim, Y. S., Wagner, R. K., & Lopez, D. (2012). Developmental relations between reading fluency and reading comprehension: A longitudinal study from Grade 1 to Grade 2. *Journal of Experimental Child Psychology,* 113(1), 93–111. doi:10.1016/j.jecp.2012.03.002

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly,* 45, 232–253.

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices*. Journal of Educational Psychology,* 95, 3–21*.*

Liu, T., Chen, W., Liu, C. H., & Fu, X. (2012). Benefits and costs of uniqueness in multiple object tracking: The role of object complexity. *Vision Research, 66*, 31–38. https://doi.org/10.1016/j.visres.2012.06.009

Mason, M. (1980). Reading ability and the encoding of item and location information. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 89.

McMahon, W. (2000). *The impact of human capital on non-market outcomes and feedback on economic development*. Paris, France: Organization for Economic Co-operation.

McMahon, W. (2002). *Education and development: Measuring the social benefits*. Oxford, UK: Oxford University Press.

Moore, A.-M., Gove, A., & Tietjen, K. (2017). Great expectations: A framework for assessing and understanding key factors affecting student learning of foundational reading skills. In A. Gove, A. Mora, & P. McCardle (Eds.), Progress toward a literate world: Early reading interventions in low-income countries, New Directions for Child and Adolescent Development, 155, 13–30.

Nag, S., & Snowling, M. J. (2011). Reading in an alphasyllabary: Implications for a language universal theory of learning to read. *Scientific Studies of Reading, 16*(5)*,* 404–423. https://doi.org/10.1080/10888438.2011.576352

Nakamura, P. & de Hoop, T. (2014). Facilitating reading acquisition in multilingual environments in India (FRAME-India): Final Report. American Institutes for Research. Available from http://www.air.org/sites/default/files/downloads/report/FRAME_Final%20Report_Final.pdf

Pae, H. K., & Sevcik, R. A. (2011). The role of verbal working memory in second language reading fluency and comprehension: A comparison of English and Korean*. International Electronic Journal of Elementary Education, 4(1),* 47–65.

Petscher, Y., & Kim, Y. S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology,* 49(1), 107–129. doi:10.1016/j.jsp.2010.09.004

Piper, B., Schroeder, L., & Trudell, B. (2016). Oral reading fluency and comprehension in Kenya: Reading acquisition in a multilingual environment. *Journal of Research in Reading, 39*(2), 133–152.

Pretorius, E. J., & Spaull, N. (2016). Exploring relationships between oral reading fluency and reading comprehension amongst English second language readers in South Africa. *Reading and Writing, 29*(7), 1449–1471. https://doi.org/10.1007/s11145-016-9645-9

Rasinski, T. V. (2011). Teaching reading fluency. In Rasinski, T. V. (Ed.), *Rebuilding the foundation: Effective reading instruction for 21st century literacy*. Bloomington, IN: Solution Tree Press.

RTI International. (2015). *Early Grade Reading Assessment (EGRA) toolkit, Second Edition.* Washington, DC: United States Agency for International Development.

Room to Read. (2016). *Khmer fluency benchmarking report*. San Francisco, CA: Author.

Samuels, S. J. (2002). Reading fluency: Its development and assessment. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 166–183). Newark, DE: International Reading Association, Inc.

Samuels, S. J. (2006). Toward a model of reading fluency. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 24–46).

Sen, B. (1997). Health and poverty in Bangladesh. *World Health*, *50*(5), 28–32.

Shen, H. H., & Jiang, X. (2013). Character reading fluency, word segmentation accuracy, and reading comprehension in L2 Chinese. *Reading in a Foreign Language, 25(1),* 1–25.

Stage, S. A. & Jacobson, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review (30),* 407–419.

Stanovich, K. (2000). Progress in understanding reading: Scientific foundations and new frontiers. New York: The Guilford Press.

Sweet, A. P., & Snow, C. E. (2003). *Rethinking reading comprehension.* New York: Guilford Press.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2005). *Education for all: The quality imperative.* Global Monitoring Report. Paris, France: Author.

UNESCO. (2015). *Education for all 2000–2015: Achievements and challenges.* Global Monitoring Report. Paris, France: Author.

Winskel, H. (2014). Introduction. In H. Winskel & P. Padakannaya (Eds.), *South and Southern Asian psycholinguistics*. Cambridge, UK: Cambridge University Press.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific studies of reading, 5*(3), 211–239.

Wolfe, B. L., & Haveman, R. H. (2002, June). Social and nonmarket benefits from education in an advanced economy. In *Conference series-Federal Reserve Bank of Boston* (Vol. 47, pp. 97–131). Boston, MA: Federal Reserve Bank of Boston.

Yanagizawa-Drott, D. (2012). *Propaganda and conflict: Theory and evidence from the Rwandan genocide*. CID Working Paper No. 257. Cambridge, MA: Harvard University Center for International Development.

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service.