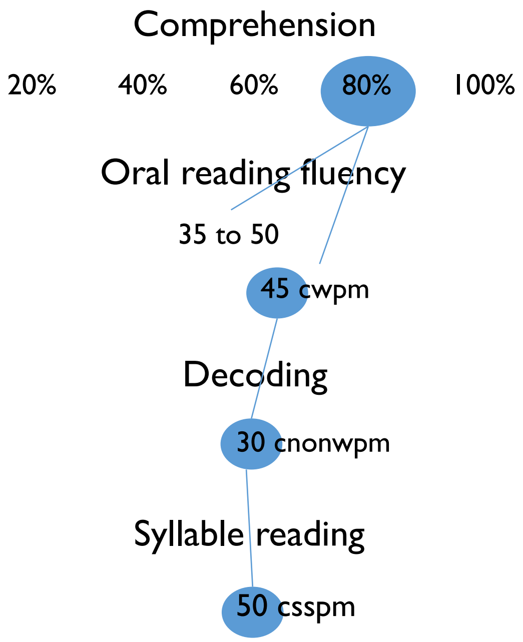
USAID
FROM THE AMERICAN PEOPLE

Some Basics for Leading Benchmarking Work Using Data from the Early Grade Reading Assessment

1. The Process

Benchmarking should rely on actual data on student performance in specific reading skill areas. The underlying relationships between the reading skill areas—in terms of both the research on how students learn to read in alphabetic languages, and the statistical relationships that have consistently been demonstrated across scores of EGRA applications—are what make it possible to use EGRA data to set benchmarks.



Step 1: Begin by discussing the level of reading comprehension that is acceptable as demonstrating full understanding of a given text. Most countries have settled on 80% or higher (4 or more correct responses out of 5 questions) as the desirable level of comprehension.

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 should be 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).

Step 3: With an ORF benchmark defined, the relationship between ORF and decoding (nonword reading) makes it possible to identify the average rate of nonword reading that corresponds to the given level of ORF.

Step 4: The process then proceeds in the same manner for each subsequent skill area.

Some tips regarding this process:

- A minimum, yet still adequate, approach to benchmarking would include two skill areas: reading comprehension and oral reading fluency.
- Going beyond those two to develop benchmarks for other skill areas can be useful, especially in countries where all of the EGRA-measured skills are poorly developed (so that progress can be detected in students' development of more basic skills).
- Syllable reading (especially when syllables are important components of words, such as in Bantu languages) is a good skill area to include.
- If syllable reading was not tested, letter sound recognition, not letter naming, should be used. An exception would be in a language like Bahasa Indonesia which is totally transparent, and therefore in which letter names and sounds are essentially the same.

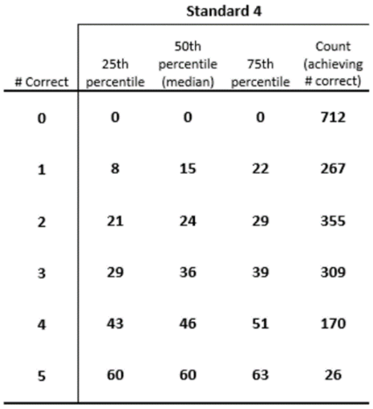
2. The Data

A good benchmarking exercise is quite data intensive. In fact, one of the added benefits of doing this exercise in a country is that the participants get to engage the EGRA results in a much deeper way than they normally would, leading to a richer understanding of what is happening in the country in terms of skill development.

The data needed to do benchmarking include:

- A table (like the one to the right) showing the range of reading fluency scores obtained by students achieving each level of reading comprehension. This makes it possible for participants to complete steps 1 and 2 in the benchmarking process.

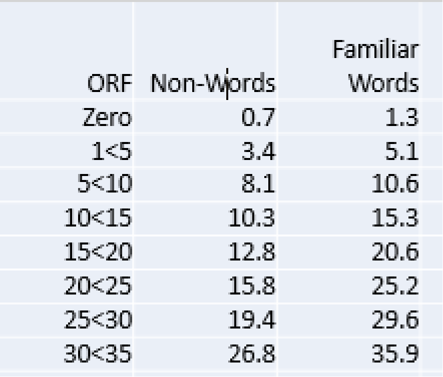
  A graphic way to depict this same information is a set of "box and whisker" plots showing the distribution of ORF scores for each level or reading comprehension.



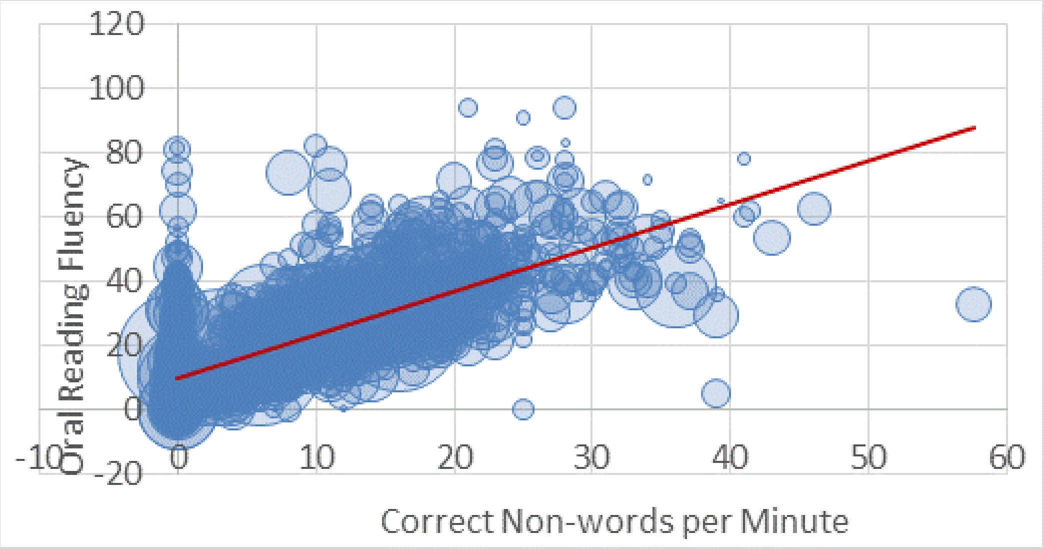
	Standard 4			
# Correct	25th percentile	50th percentile (median)	75th percentile	Count (achieving # correct)
0	0	0	0	712
1	8	15	22	267
2	21	24	29	355
3	29	36	39	309
4	43	46	51	170
5	60	60	63	26

- A table that shows the average scores on each other subtask that correspond to different levels of oral reading fluency (as shown here) is what enables participants to connect the ORF benchmark to desirable levels of performance in other skill areas.

  A graphic way to show this same information is to use a scatter plot (below)—for example, of ORF x nonword decoding, with the best fit line drawn in so that workshop participants can match a given level of ORF to the average corresponding level of nonword decoding.



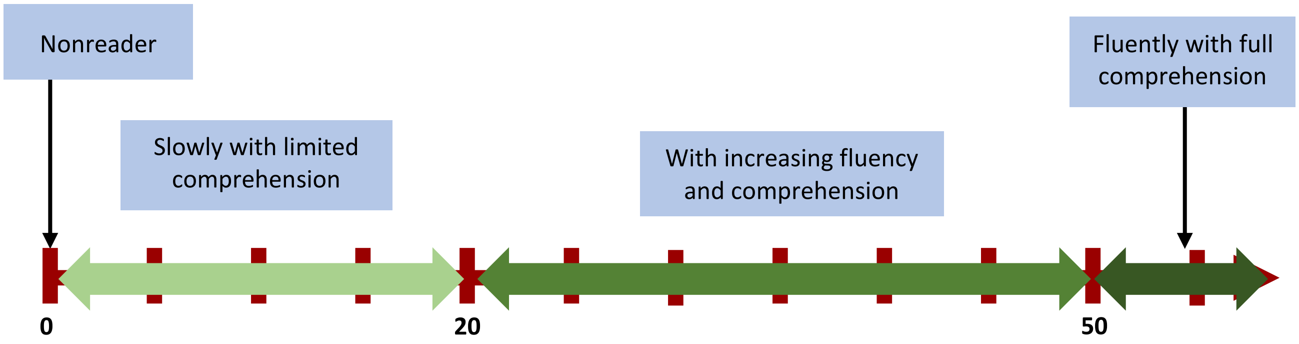
ORF	Non-Words	Familiar Words
Zero	0.7	1.3
1<5	3.4	5.1
5<10	8.1	10.6
10<15	10.3	15.3
15<20	12.8	20.6
20<25	15.8	25.2
25<30	19.4	29.6
30<35	26.8	35.9



- For determining the percentage of students meeting the benchmark (in the year for which the EGRA data are available), a cumulative distribution graph or table makes it possible for participants to "look up" the percentage of students, for example, achieving 45 cwpm or higher.

3. Performance Levels

Some countries are interested in establishing performance points that capture stages of skill development that are below the desired level of achievement defined by the benchmark. For example, the benchmark for reading fluency may be defined as 50 cwpm, representing students who are reading fluently and with full (or almost full) comprehension. Students who score zero are those who are not reading. In between zero and 50 cwpm exist different levels of reading ability that in fact may correspond to stages of literacy acquisition. Setting multiple performance levels makes it possible to determine what percentages of children are at each of those stages of development of their reading skill.


Nonreader
Slowly with limited comprehension
With increasing fluency and comprehension
Fluently with full comprehension
0
20
50

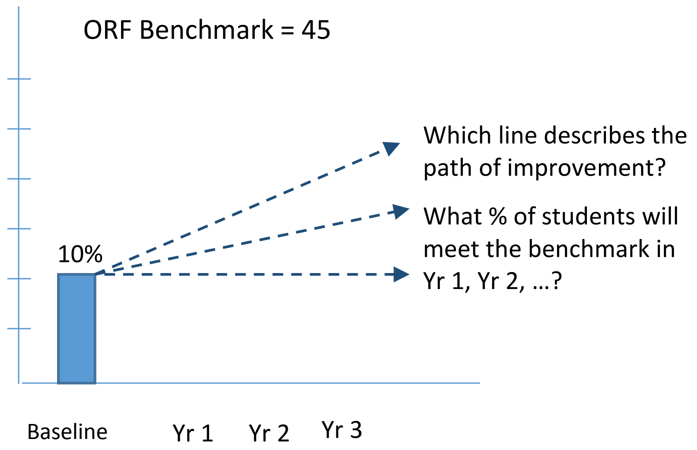
As illustrated above, it is possible to create two other performance levels below the benchmark for reading fluently with full comprehension set at 50 cwpm. Data describing how reading fluency and comprehension scores are distributed (e.g., using a two-way distribution table) inform where to place another level of reading achievement in between zero and 50 cwpm. Thus, two other performance levels are created: students who score above zero and up to 20 cwpm are said to be reading slowly with limited comprehension, and those scoring above 20 and up to 50 cwpm can be said to be reading with increasing fluency and comprehension. The performance levels in this example are from the benchmarking work done in Ethiopia in early 2015. Such intermediate performance levels in other contexts could, of course, be given other labels.

An alternative approach to setting performance levels (as was the case in Pakistan) would be to establish a range of ORF scores that are defined as "meeting expectations": 60 to 90 cwpm. Students scoring above 90 cwpm would be considered to be "exceeding expectations." Those scoring below 60 would be "not meeting expectations."

4. Moving Beyond Benchmarks to Targets

One of the main purposes of setting benchmarks is to establish the means to evaluate and measure progress in improving reading outcomes. In fact, one of the more interesting challenges in working with ministry colleagues to set benchmarks arises during discussions of the prospects for future improvement in student performance relative to those benchmarks. To set targets for future improvement, benchmarks can be used in the following way.

Once a benchmark has been set, say for oral reading fluency, it is useful to employ the existing data to determine the percentage of students presently meeting that benchmark. The challenge arises when assumptions have to be made about how things will improve—that is, to estimate the percentages of students who will meet the benchmark in future years (as illustrated here).



If data are available from a reading intervention in the country, then the amount of improvement achieved by that program provides a useful starting point for estimating future targets.

If data from an intervention are not available, but EGRA results from more than one year are, then the prevailing pattern of change over time can be used to begin discussing how that pattern may evolve in future years.

If only one year of reading results is available, then the task is less data-driven and more a dialogue about how much improvement can be expected. Data from other countries' programs that have had demonstrated impact could inform that dialogue. Additionally, if EGRA data from a given year are available for two successive grades (say, grades 1 and 2), then the "intergrade" difference is a good means for estimating how much improvement to expect.

The intergrade difference represents the amount of progress students make given an additional year in school (under preexisting conditions). For example, a successful intervention could aim to improve performance in grade 1 by as much as the intergrade difference between grades 1 and 2; or put differently, to increase student performance by as much as an additional year of schooling.

The value of setting targets is that if performance is initially low—say, very few students meeting the benchmark—there often emerges a tendency to want to lower the benchmark (so performance does not look as bad). It is better to have a benchmark that is genuinely meaningful in terms of the skill level achieved (e.g., oral reading that is fluent enough to enable students to comprehend what they are reading). Therefore, instead of lowering the benchmark, a compromise is to have modest targets for the percentage of children expected to meet the benchmark moving forward. Examples of benchmarks and targets from Jordan are shown in the table below.

		Oral Reading Fluency	Nonword Decoding
Benchmark		46 cwpm	23 cnonwpm
% of students meeting the benchmark	2014 actual	7.5%	5.3%
	5-year target	35%	31%

Participants estimated that the percentage of students meeting the benchmarks for these two skill areas would increase from 7.5% to 35% and from 5.3% to 31% over the course of the next five years.

5. Some Things to Remember

Having facilitated benchmarking exercises in nine countries, the Education Data for Decision Making (EdData II) project team has learned some useful lessons, which are summarized here.

- **Supply the data.** The process requires a fair amount of data. Preparing the right data tools—graphs, tables, forms to be filled out—and carefully labeling those tools to correspond to the different steps in the process help greatly with facilitating the running of a benchmarking workshop.

- **Match the data to the task.** A balance needs to be struck between too much and too little data. When a lot of data are available (from more than one year of EGRA, for multiple grades, for an intervention as well as from national surveys), be sure to have participants working only with the sets of data that correspond to the task at hand. Do not dump everything on them at once.

- **Work across grade levels.** For working with more than one grade—e.g., for grades 1 through 3—it is best to work in each skill area across grades. For example, when setting a benchmark for ORF, set it for the highest grade for which data are available and then work to set the benchmarks for the other two grades based on that. Then move on to do the same in another skill area.

- **Have multiple small groups work simultaneously.** It is useful to have more than one group working in parallel with the data to set a benchmark. When groups arrive at different suggested benchmarks, the facilitated dialogue that ensues is usually quite fruitful. And that dialogue illustrates that even when everyone is using data, there is room for interpretation and negotiation about what constitutes a reasonable benchmark for a given country and language.

- **Encourage discussion.** Similarly, the discussion, and often debate, about what targets should be set for future improvement brings to the surface everyone's assumptions about how the system is going to improve over time. For example, when looking at the results of a pilot interventions in Malawi and Liberia as the bases for determining future targets, participants had a lively discussion about whether one could assume that the conditions created in a pilot (which led to the results) could be expected to be implemented on a national scale (and what it would take for the ministry and its partners to achieve that).

- **Limit the number of benchmarks.** There is often a tendency to want to set benchmarks for every skill area. Limiting the number of skill areas to no more than four is highly recommended: reading comprehension, oral reading fluency, and two others (nonword decoding and syllable reading or letter sound identification).

- **Consider how to institutionalize the decisions.** It is necessary to engage participants in determining how benchmarks they develop could become official. Even if the benchmarks are not made official, they should be used to summarize reading performance the next time early grade reading is assessed. This was the case most recently in the Philippines where, even though the benchmarks were not officially adopted, comparison of the percentages of students meeting benchmarks in 2014 and 2015 helped the Department of Education evaluate the extent to which progress was being made.