



# EGRA FRAMEWORK

## Toolkit for the Early Grade Reading Assessment Adapted for Zambia

February 2016

Ministry of General Education (MOGE)  
and  
Examinations Council of Zambia (ECZ)





# EGRA FRAMEWORK

## Toolkit for the Early Grade Reading Assessment Adapted for Zambia

Ministry of General Education

and

Examinations Council of Zambia



This manual was made possible by the support of the American people through the United States Agency for International Development (USAID). USAID funding for the *Early Grade Reading Assessment Toolkit, Adapted for Zambia*, was provided through an Education Data for Decision Making (EdData II) task order, Data Collection Services for the USAID/Zambia Education Project, Contract No. AID-611-M-14-00002 (RTI International Task 28).

The authors' views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

Education Office United States Agency for International  
Development (USAID)/Zambia  
U.S. Embassy  
Subdivision 694 / Stand 100  
P.O. Box 32481  
Kabulonga District, Ibex Hill Road  
Lusaka, District  
Zambia

RTI International  
3040 Cornwallis Road  
Post Office Box 12194  
Research Triangle Park, NC 27709-2194  
USA

Cover photo: EGRA data collection in Zambia via electronic tablet. (Credit: RTI staff)



# CONTENTS

	PAGE
List of Exhibits .....	vi
Abbreviations.....	viii
Glossary of Terms.....	x
Reading-Related Terminology .....	x
Statistical Terms .....	xi
Methodological Terms.....	xii
1 Introduction.....	1
1.1 Why Do We Need Early Grade Reading Assessments (EGRAs)? .....	1
1.1.1 Why Assess <b>Reading</b> ? .....	2
1.1.2 Why Assess <b>Early</b> ? .....	2
1.1.3 Why Assess <b>Orally</b> ? .....	4
1.1.4 EGRA's Place Among Assessment Options .....	5
1.2 Development of the EGRA Instrument.....	6
1.3 The Instrument in Action .....	7
1.4 EGRA's Presence in Zambia .....	8
2 Purpose and Uses of EGRA .....	10
2.1 History and Overview .....	10
2.2 EGRA as a System Diagnostic .....	11
3 Conceptual Framework and Research Foundations.....	13
3.1 Summary of Skills Necessary for Successful Reading .....	13
3.2 Phonological Awareness .....	14
3.2.1 Description .....	14
3.3 The Alphabetic Principle, Phonics, and Decoding .....	15
3.3.1 Description .....	15
3.3.2 Measures of Alphabet Knowledge and Decoding Skills .....	16
3.4 Vocabulary and Oral Language .....	16
3.4.1 Description .....	16
3.4.2 Measures of Vocabulary .....	17
3.5 Fluency .....	17
3.5.1 Description .....	17
3.5.2 Measures of Fluency.....	18
3.6 Comprehension .....	19
3.6.1 Description .....	19

	3.6.2	Measures of Reading Comprehension .....	19
4		EGRA Instrument Design: Adaptation Development and Adaptation Modification .....	20
	4.1	Adaptation Workshop .....	20
	4.1.1	Overview of Workshop Planning Considerations .....	21
	4.1.2	Who Participates? .....	22
	4.1.3	What Materials Are Needed? .....	22
	4.2	Review of the Zambian Instrument Components .....	23
	4.2.1	Listening Comprehension .....	24
	4.2.2	Letter Sound Identification .....	27
	4.2.3	Nonword Reading .....	31
	4.2.4	Oral Reading Passage with Comprehension .....	33
	4.2.5	Orientation to Print .....	34
	4.2.6	English Vocabulary .....	35
	4.3	Translation and Other Language Considerations .....	37
	4.3.1	Translation vs. Adaptation .....	37
	4.3.2	Cross-Language Comparisons: Preparations and Considerations .....	38
	4.4	Using Same-Language Instruments Across Multiple Applications: Creation of Equivalent Test Forms .....	40
	4.5	Best Practices .....	40
5		Using Electronic Data Collection .....	42
	5.1	Cautions and Limitations to Electronic Data Collection .....	43
	5.2	Data Collection Software .....	43
	5.3	Considerations for Hardware Selection and Purchasing .....	44
	5.4	Supplies Needed for Electronic Data Collection and Training .....	44
6		EGRA Assessor Training .....	45
	6.1	Recruitment of Training Participants .....	45
	6.2	Planning the Training Event .....	47
	6.3	Components of Assessor Training .....	48
	6.4	Training Methods and Activities .....	48
	6.5	School Visits .....	49
	6.6	Assessor-Trainee Evaluation Process .....	51
	6.7	Measuring Assessors' Accuracy .....	52
7		Field Data Collection: Pilot Test and Full Study .....	55
	7.1	Conducting a Pilot EGRA .....	55
	7.1.1	Pilot Study Data and Sample Requirements .....	56
	7.1.2	Establishing Test Validity and Reliability .....	57
	7.1.3	Considerations Regarding the Timing of the Pilot Test .....	59
	7.2	Field Data Collection Procedures for the Full Studies .....	60
	7.3	Selecting Students .....	62
	7.3.1	Student Sampling Option 1: Random Number Table .....	62

7.3.2	Student Sampling Option 2: Interval Sampling .....	63
7.4	End of the Assessment Day: Wrapping Up.....	65
7.5	Uploading Data Collected in the Field.....	65
8	Preparation of EGRA Data.....	67
8.1	Data Cleaning.....	67
8.2	Processing of EGRA Subtasks .....	68
8.2.1	<prefix>_ .....	69
8.2.2	<suffix>.....	69
8.3	Timed Subtasks.....	70
8.4	Untimed Subtasks .....	71
8.5	Statistical Equating.....	72
9	Data Analysis and Reporting.....	75
9.1	Descriptive Statistics (Non-inferential) .....	75
9.2	Types of Regression Analysis.....	75
9.3	Reporting Data Analysis.....	76
10	Using Results to Inform Action.....	78
10.1	Setting Country-Specific Benchmarks.....	78
10.1.1	What Are Benchmarks? .....	78
10.1.2	Criteria for Establishing Benchmarks.....	79
10.1.3	A Process for Setting Benchmarks .....	81
	Bibliography.....	83
	Annex A: Recommendations and Considerations for Cross-Language Comparisons .....	95
	Annex B: Sample Assessor Training Agenda .....	98
	Annex C: Data Analysis and Statistical Guidance for Measuring Assessors' Accuracy.....	100
	Annex D: Sample Codebook.....	104

# LIST OF EXHIBITS

	PAGE
Exhibit 1. Reading trajectories of low and middle readers: Reading fluency (measured in correct words per minute) .....	3
Exhibit 2. Student words per minute scores, grades 1 and 2 .....	4
Exhibit 3. Different types of assessments: A continuum .....	5
Exhibit 4. Map of EGRA administrations .....	7
Exhibit 5. Worldwide application of the EGRA instrument: Number of countries, by year .....	8
Exhibit 6. The continuous cycle of improving student learning .....	11
Exhibit 7. Differences between EGRA adaptation development and adaptation modification .....	21
Exhibit 8. Sample agenda: EGRA adaptation development or adaptation modification workshop .....	23
Exhibit 9. Review of Zambian instrument components .....	24
Exhibit 10. Sample: Listening comprehension (English) .....	26
Exhibit 11. Sample: Listening comprehension (Chinyanja) .....	27
Exhibit 12. Sample: Assessor protocol, letter sound identification (Icibemba language, Zambia) .....	30
Exhibit 13. Sample: Student stimulus sheet, letter sound identification (Icibemba language, Zambia) .....	31
Exhibit 14. Sample: Nonword reading (Icibemba language, Zambia) .....	32
Exhibit 15. Sample: Oral reading passage with reading comprehension (Luvale language, Zambia) .....	34
Exhibit 16. Sample: Orientation to print (Kikaonde language, Zambia) .....	35
Exhibit 17. Sample: English vocabulary knowledge (instructions in Lunda language) .....	37
Exhibit 18. Frame from video used for assessment .....	53
Exhibit 19. Differences between EGRA pilot test and full data collection .....	56
Exhibit 20. Determinants of the sampling groups .....	63

Exhibit 21. Data cleaning checklist .....	68
Exhibit 22. EGRA subtask variable nomenclature and names of the timed score variables .....	69
Exhibit 23. Suffix nomenclature for the item and score variables .....	70
Exhibit 24. Sample counterbalanced design.....	73

# ABBREVIATIONS

ASER	Pratham’s Annual Status of Education Report assessment
CONFEMEN	Conférence des Ministres de l’Éducation des Pays ayant le Français en Partage
CTT	classical test theory
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
DFID	UK Department for International Development
ECZ	Examinations Council of Zambia
EDC	Education Development Center, Inc.
EFA	Education for All
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
EMIS	education management information system
GPC	grapheme-phoneme correspondence
GPS	global positioning system
ICC	intra-class correlation coefficient
ID	identification
IPA	International Phonetic Alphabet
IRR	interrater reliability
IRT	item response theory
ISO	International Organization for Standardization
LCD	liquid-crystal display
LLC	limited-liability company
LLECE	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación
LQAS	lot quality assurance sampling
MDG	Millennium Development Goal
MOGE	Ministry of General Education (Zambia) <sup>1</sup>
NAS	National Assessment Survey (Zambia)
NCFL	National Center for Family Literacy
NICHD	US National Institute for Child Health and Human Development
OLS	ordinary least squares
ORF	oral reading fluency
PASEC	Programme d’Analyse des Systèmes Educatifs de la CONFEMEN

---

<sup>1</sup> Previously known as the Ministry of Education, Science, Vocational Training, and Early Education (MESVTEE).

PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment,
PPS	probability proportional to size
PRIMR	Primary Math and Reading Initiative, Kenya
RTI	RTI International (registered trademark and trade name of Research Triangle Institute)
RTS	Read to Succeed
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
TIMSS	Trends in International Mathematics and Science Study
TTL	Time to Learn
UNDP	United Nations Development Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
USAID	United States Agency for International Development

# GLOSSARY OF TERMS

## Reading-Related Terminology

**Alphabetic knowledge/process.** Familiarity with the alphabet and with the principle that written letters systematically represent sounds that can be blended into meaningful words.

**Blend.** A group of two or more consecutive consonants that begin a syllable (as *gr-* or *pl-* in English). This is different from a digraph because the letters keep their separate sounds when read.

**Derivation.** A word formed from another word or base, such as *farmer* from *farm*.

**Digraph.** A group of consecutive letters that combine to make a single sound (e.g., *ea* in *bread*, *ch* in *chin*). Some digraphs are *graphemes* (see below).

**Fluency / Automaticity.** The bridge between decoding and comprehension. Fluency is being able to read words quickly, accurately, and with expression (prosody). This skill allows readers to use more mental effort on making meaning than translating letters to sounds and forming sounds into words. At that point, readers are decoding quickly enough to be able to focus most of their effort on comprehension.

**Fluency analysis.** A measure of overall reading competence reflecting the ability to read accurately and quickly (see Fluency / Automaticity).

**Grapheme.** The most basic unit in an alphabetic written system that can change the meaning of a word. Graphemes represent phonemes. A grapheme might be composed of one or more than one letter; or of a letter with a diacritic mark (such as “*é*” vs. “*e*” in French).

**Inflected form.** A change in a base word in varying contexts to adapt to person, gender, tense, etc.

**Morpheme.** Smallest linguistic unit with meaning. Different from a word, as words can be made up of several morphemes (e.g., “unbreakable” can be divided into *un-*, *break*, and *-able*). There are **bound** and **unbound** morphemes. A word is an unbound morpheme, meaning that it can stand alone. A bound morpheme cannot stand alone (e.g., prefixes such as *un-*).

**Onset.** The first consonant or consonant cluster that precedes the vowel of a syllable; for example, *spoil* is divided into “*sp*” (the onset) and “*oil*” (the *rime*; see below).

**Orthography.** The written representation of the sounds of a language; spelling.

**Phoneme.** The smallest linguistically distinctive unit of sound that changes the meaning of a word (e.g., “*top*” and “*mop*” differ by only one phoneme, but the meaning changes).

**Phonological awareness.** Awareness that words are made of sounds; and the ability to hear, identify, and manipulate these sounds. Sounds exist at three levels of structure: syllables, *onsets* and *rimes*, and *phonemes* (see italicized terms).

**Phonics.** Instructional practices that emphasize how spellings are related to speech sounds in systematic ways.

**Rime.** The part of a syllable that consists of its vowel and any consonant sounds that come after it; for example, *spoil* is divided into “sp” (the *onset*; see above) and “oil” (the *rime*).

## Statistical Terms

**Ceiling effect.** Occurs when there is an artificial upper limit on the possible values for a variable and a large concentration of participants score at or near this limit. This is the opposite of the *floor effect* (see below). For example, if an EGRA subtask is much too easy for most children, the scores will concentrate heavily at the upper end of the allowable range, restricting the variation in scores and negatively impacting the validity of the tool itself.

**Convenience sample.** Also known as reliance on available subjects, a convenience sample is a nonprobability sample that relies on data collection from population members who are easy to reach (or conveniently available). This method does not allow for generalizations and is of limited value in social science research.

**Floor effect.** Occurs when there is an artificial lower limit on the possible values for a variable and a large concentration of participants score at or near this limit. This is the opposite of the *ceiling effect* (see above). For example, if an EGRA subtask is much too difficult for most children, the scores will concentrate heavily at the lower end of the allowable range (typically with large proportions of zero scores), restricting the variation in scores and negatively impacting the validity of the tool itself.

**Intra-class correlation coefficient (ICC).** This is a descriptive statistic that is used when data are clustered into groups. The statistic ranges from 0 and 1 and provides a measure of how closely members of a group resemble each other in certain observable characteristics. ICCs can also be used to measure consistency of measurement across observers.

From Fleiss (1981):

Kappa Statistic	Strength of Agreement
Less than 0.40	Poor
0.40 to 0.75	Intermediate to Good
Greater than 0.75	Excellent

**Kappa.** Measures the extent to which two different ratings of the same subject could have happened by chance. Kappa values range from -1.0 to 1.0. Higher values indicate lower probability of chance agreement.

**Population.** The theoretical group of subjects (individuals or units) to whom a study’s results can be generalized. The *sample* (see below) and the population share similar characteristics, and the *sample* is a part of the population of interest.

**Raw % agreement.** Measures the extent to which raters make exactly the same judgment. Due to the lack of detail provided solely by this statistic, no benchmark is possible. Ideally, raters' % agreement will be as high as possible (close to 100%) when they assess students. However, regardless of the % agreement, Kappa statistics must be referenced to understand the quality of the % agreement statistic.

**Sample.** The group of subjects (individuals or units) selected to be in a study.

**Sampling unit.** The individual members of the *sample* (see above); the unit from which data will be collected. For example, individuals or households may be the sampling unit.

**Simple random sampling.** The simplest form of probability sampling. Simple random sampling is a method in which every member of the *population* has the same probability of being selected for inclusion in the *sample* (see entries for italicized terms).

**Test reliability.** The consistency of scores a test-taker would receive on two different but equally difficult forms of the same test.

## Methodological Terms

**Attrition.** The gradual loss of subjects; often occurs in longitudinal studies when subjects drop out of the study before it is completed, for example, between the baseline and the midline.

**Content validity.** Term used to indicate the degree to which the items are representing the measurement of the intended skills.

**Control group.** Subjects who are randomly assigned not to receive treatment (intervention) and whose characteristics of interest are compared with those of a treatment group following the treatment.

**Comparable test forms.** Tests that are intended to be judged in relationship to each other and thus are designed with the same constructs, subtasks, etc.

**Comparison group.** Subjects who do not receive treatment (intervention) but are similar to the ones who receive the intervention, and whose characteristics of interest are compared to those of the treatment group following the treatment. Frequently selected based on similarity of certain characteristics with the treatment group (also called "matched comparison group").

**Equated test forms.** Refers to test forms that have been adjusted by a statistical process in order to make scores comparable.

**Equivalent test forms.** Tests that are intended to be of equal difficulty (and thus directly substitutable for one another).

**Face validity.** Term used to indicate the extent to which a test overall is viewed as covering the concepts its designers intended to measure.

# 1 INTRODUCTION

## 1.1 Why Do We Need Early Grade Reading Assessments (EGRAs)?

Countries around the world have boosted primary school enrollment to historically unprecedented rates. Thanks to the targeted efforts of the United Nations' Education for All (EFA) campaign and the Millennium Development Goals (MDGs) that were slated for achievement by 2015, the world has seen dramatic improvements in primary school enrollment rates; in some places they are now nearly the same rates as in high-income countries. The net enrollment rate for primary school in developing regions has reached an estimated 91 percent in 2015, up from 83 percent in 2000; and the number of out-of-school children of primary school age worldwide has fallen by almost half in the same time frame (United Nations, 2015).

Data on results from low-income countries that have participated in various international assessments—including tests administered in grades 1 through 3—are now available for comparison on the World Bank's online EdStats Dashboard pages on learning outcomes (World Bank, 2015). However, the evidence still indicates that while enrollment has increased, average student learning in most low-income countries remains quite low. The World Bank recently summarized the situation thus: "There is broad consensus among the international community that the achievement of the education Millennium Development Goal (MDG) requires improvements in learning outcomes" (World Bank, 2015b); and Quality Education was adopted globally as Goal 4 of the post-2015 Sustainable Development Goals (United Nations Development Programme [UNDP], 2015). The importance of education quality for national economic development is another area of broad agreement: "Recent research reveals that it is learning rather than years of schooling that contributes to a country's economic growth: A 10 percent increase in the share of students reaching basic literacy translates into an annual growth rate that is 0.3 percentage points higher than it would otherwise be for that country" (Hanushek & Woessman, 2009, as cited in Gove & Wetterberg, 2011, pp. 1–2).

At the time the first edition of this toolkit was written in 2009, the most commonly used assessments were able to reveal what low-income country students did not know, but could not ascertain what they did know, often because they scored so poorly that the test could not pinpoint their location on the knowledge continuum. Furthermore, most national and international assessments were historically administered as paper-and-pencil tests to students in grade 4 and above (that is, they assumed students could read and write). It was not always possible to tell from the results of these tests whether students scored poorly because they lacked the knowledge tested by the assessments, or because they lacked basic reading and comprehension skills. Since 2010, a turn toward reading-skill assessments in the early grades—due in large part to the influence of the United States Agency for International Development (USAID) and the World Bank—marks a change in awareness among international education

researchers and stakeholders regarding the need for more empirical information about young children’s ability to read with comprehension.

The ability to read and comprehend a simple text is one of the most fundamental skills a child can learn. Without basic literacy there is little chance that a child can escape the intergenerational cycle of poverty. Yet in many countries, students enrolled in school for as many as six years are unable to read and understand a simple text. Evidence indicates that learning to read both early and at a sufficient rate (with comprehension) is essential for learning to read well.

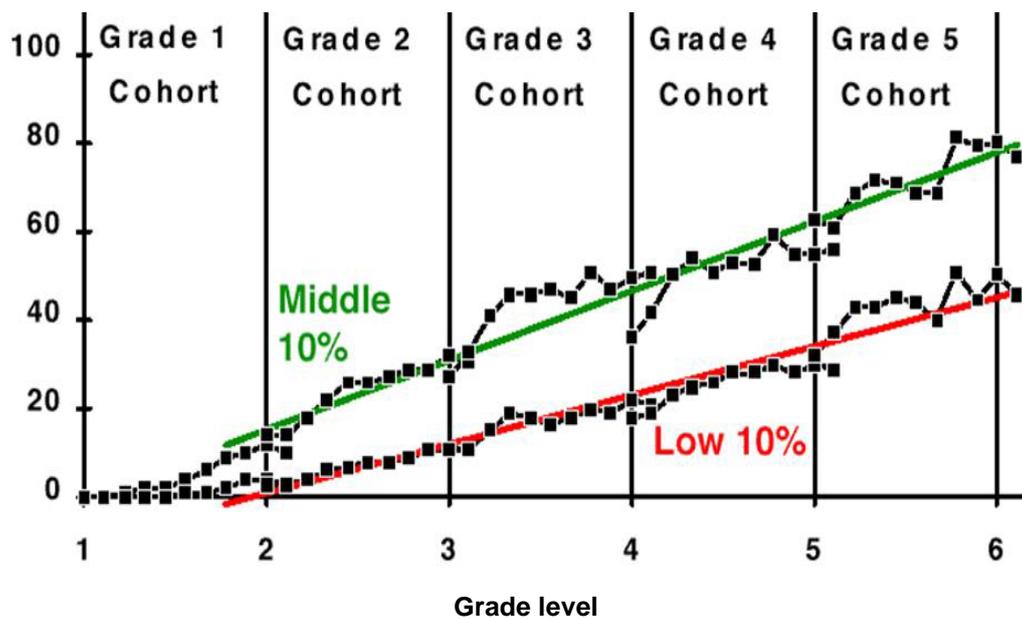
### 1.1.1 Why Assess Reading?

Basic literacy is the foundation children need to be successful in all other areas of education. Children first need to “learn to read” so that they can “read to learn.” That is, as children pass through the grade levels, more and more academic content is transmitted to them through text, and their ability to acquire new knowledge and skills depends largely on their ability to read and extract meaning from text. For example, math is an important skill, but using a math book requires the ability to read. Students are also increasingly required to demonstrate their learning through writing, a skill integrally tied to reading. Moreover, a low level of literacy severely constrains a person’s capacity for self-guided and lifelong learning that is so important beyond the classroom walls into the world of adult responsibilities.

### 1.1.2 Why Assess Early?

Acquiring literacy becomes more difficult as students grow older; children who do not learn to read in the first few grades are more likely to repeat grades and eventually drop out. That is, if strong foundational skills are not acquired early on, gaps in learning outcomes (between the “haves” and the “have-nots”) grow larger over time (see **Exhibit 1** as well as Adolf et al., 2010; Daniel et al., 2006; Darney, Reinke, Herman, Stormont, & Jalongo, 2013; Scanlon, Gelzheiser, Vellutino, Schatschneider, & Sweeney, 2008; Torgesen, 2002). The common metaphor of “the rich get rich and the poor get poorer” is often quoted in discussions of the disparities that occur between fluent and nonfluent readers for children who are unable to acquire reading and comprehension skills in the early grades (Gove & Wetterberg, 2011).

**Exhibit 1. Reading trajectories of low and middle readers: Reading fluency (measured in correct words per minute)**



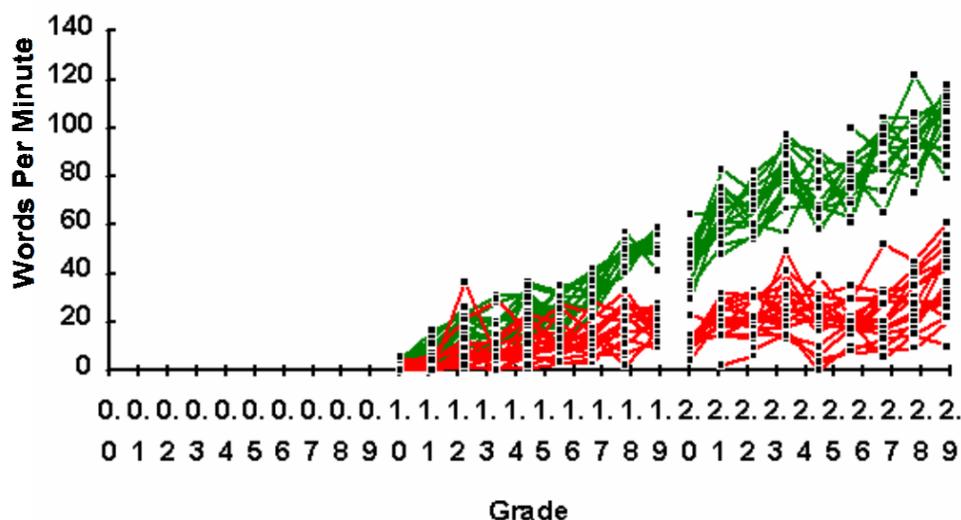
Source: Good, Simmons, & Smith, 1998, Figure 1.

Unlike many skills such as walking and speaking, the ability to read is not acquired naturally without instruction. Studies suggest that without quality instruction, a child who reads poorly in the early grades will continue to read poorly in the upper grades, and will require more and more instructional intervention in order to “catch up” (Juel, 1988).

**Exhibit 2** documents the trajectory of student performance on oral reading fluency for a group of students during grades 1 and 2 in the United States among students who did not receive additional instruction for reading improvement. The cluster of lines in the upper part of the left side of the graph shows monthly results for students who could read at least 40 words per minute at the end of first grade, while the cluster of lines at the bottom shows the results for students who read fewer than 40 words per minute at the end of first grade. (Each unit on the horizontal axis represents a month in the school year.)

As can be seen in Exhibit 2, the gap between more proficient and less proficient readers increases dramatically by the end of second grade (right side of graph). In the absence of timely intervention or remediation, this initial gap in reading acquisition is likely to widen over time and become increasingly difficult to bridge.

## Exhibit 2. Student words per minute scores, grades 1 and 2



Source: Good, Simmons, & Smith, 1998, Figure 2 (grade 1) and Figure 3 (grade 2).

Note: Numbers on the horizontal axis refer to the grade (top row) and month (bottom row).

The more children struggle at school, the greater the risk they will become discouraged and drop out, forfeiting any potential benefits that education would afford them later in life. In contrast, the more and better they learn, the longer they tend to stay in school (Patrinos & Velez, 2009). One study found that the strongest predictor of primary school completion in Senegal was the child's level of success in second grade (Glick & Sahn, 2010). Whether for an individual child or for a whole educational system, it is more efficient to address a reading deficit in the early grades than later.

### 1.1.3 Why Assess Orally?

Traditional paper-based tests require that children already have acquired basic reading fluency and comprehension skills. If they have not—i.e., if they are unable to read the question or write the answer—the results will suffer from a floor effect with a high percentage of zero scores. In those cases, the paper-based test tells us only what the children do not know, but not what they do know or where they are along the developmental path.

In many countries, students must pass a national “exit” examination at the end of grade 6 in order to earn their primary education completion certificate and/or to enter secondary school (Braun & Kanjee, 2006). Furthermore, international assessments through the Progress in International Reading Literacy Study, or PIRLS (given to fourth graders) and Programme for International Student Assessment, or PISA (given to 15-year-olds) are administered in numerous (mostly higher income) countries around the world.<sup>2</sup> In both kinds of assessments, students are generally asked to read several short passages and to answer multiple-choice questions. If the students' reading and comprehension skills are insufficient to understand the test, they will fail the assessment—but the resulting data will not reveal why they failed. Did the

<sup>2</sup> Zambia is one of seven countries participating in the PISA for Development project which launched in February 2014.

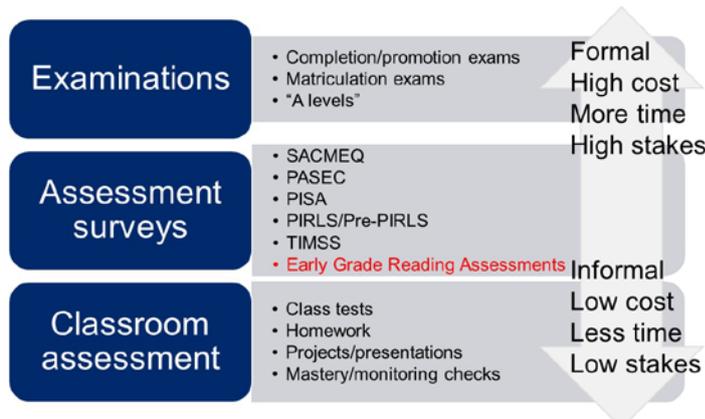
students not have the knowledge to answer the questions, or were they just unable to read the questions?

Reading fluency and comprehension are relatively higher-order skills in the reading acquisition process, and they build upon several lower-order, foundational skills such as phonological awareness, alphabet knowledge, decoding, vocabulary, etc., which can be detected through an oral assessment. An oral assessment therefore can give us more information about what they actually do know and where they are in the reading acquisition process early on. Oral assessments can also help reveal early growth over time—that is, changes that are not yet detectable on a paper-based test but that nonetheless constitute progress toward reading acquisition.

### 1.1.4 EGRA’s Place Among Assessment Options

To explain where EGRA fits in the landscape of assessment options, it is useful to place different types of assessments on a continuum (as displayed in **Exhibit 3**). The continuum is broken into three broad categories: examinations, assessment surveys, and classroom assessments. Kanjee (2009) defines examinations as processes used for testing the qualifications of candidates (e.g., quarterly exams, promotion exams, and matriculation exams). These

**Exhibit 3. Different types of assessments: A continuum**



Source: Adapted from Kanjee (2009).

are typically longer, more formal assessments that are administered to all students (thus making them more time-intensive and more costly). At the other end of the spectrum are classroom assessments, which are defined as measures used to obtain evidence on knowledge, skills, and attitudes of individual learners for the purpose of improving teaching and learning (Kanjee, 2009). These more informal assessments often come in the form of classroom tests, homework assignments, and projects/presentations. By design,

classroom assessments are intended to be cheaper, to take less time, and to involve lower stakes (particularly when compared with examinations).

Assessment surveys are designed with the explicit purpose of obtaining information on the performance of students, as well as on education systems as a whole. In addition to the PIRLS and PISA, there are many other international and regional assessments that fit into this category, such as those carried out by the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), the Programme d'Analyse des Systèmes Educatifs de la CONFEMEN<sup>3</sup> (PASEC), the Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), and the Trends in International Mathematics and Science Study (TIMSS). Since the assessments associated with these programs are intended to measure trends in

<sup>3</sup> CONFEMEN: Conférence des Ministres de l'Éducation des Pays ayant le Français en Partage.

literacy achievement for cross-country comparisons, they require long-term development processes, local language complications, and complex scaling/scoring procedures. Additionally, every one of these assessments requires basic reading ability (i.e., the assessment is based on passage reading), which limits the value and appropriateness for measuring early grade reading skills in developing countries (due to major floor effects). In recent years, new early grade reading assessments (e.g., Pratham’s Annual Status of Education Report [ASER] assessment, World Vision’s Functional Literacy Assessment Tool [FLAT]<sup>4</sup> assessment) have been developed in order to fill this gap. These individually administered assessments are touted as being “smaller, quicker, cheaper” as compared with international tests (Wagner, 2011).

## 1.2 Development of the EGRA Instrument

In the context of these questions about student learning and continued investment in education for all, departments of education and development professionals at the World Bank, USAID, and other institutions called for the creation of simple, effective, and low-cost measures of student learning outcomes (Abadzi, 2006; Center for Global Development, 2006; Chabbott, 2006; World Bank: Independent Evaluation Group, 2006). Some analysts have even advocated for the establishment of a global learning standard or goal, in addition to the original Education for All and Millennium Development Goals (Filmer, Hasan, & Pritchett, 2006) and Sustainable Development Goal 4 (UNDP, 2015). Whether reading well by a certain grade could be such a goal is open to debate, but the issue of specific and simple learning measures is now on the policy agenda.

To respond to this demand, work began on the creation of an Early Grade Reading Assessment: a simple instrument that could report on the foundation levels of student learning, including assessment of the first steps students take in learning to read. In October 2006, USAID contracted RTI International through the Education Data for Decision Making (EdData II) project to develop an instrument to help USAID partner countries begin the process of measuring in a systematic way how well children in the early grades of primary school were acquiring reading skills. Ultimately, the hope was to spur more effective efforts to improve performance in these core skills by using an assessment that can easily be adapted to new contexts and languages, has a simplified scoring system, and is low stakes and less time intensive for the individuals being assessed.

Based on a review of research and existing reading tools and assessments, RTI developed a protocol for an individual oral assessment of students’ foundational reading skills. In an initial EGRA workshop hosted by USAID, the World Bank, and RTI in November 2006, cognitive scientists, early grade reading experts, research methodologists, and assessment experts reviewed the proposed instrument and provided feedback and confirmation on the protocol and validity of the approach. The workshop included contributions from more than a dozen experts from a diverse group of countries, as well as some 15 observers from institutions such as USAID, the World Bank, the William and Flora Hewlett Foundation, George Washington University, the South Africa Ministry of Education, and Plan International, among others.

---

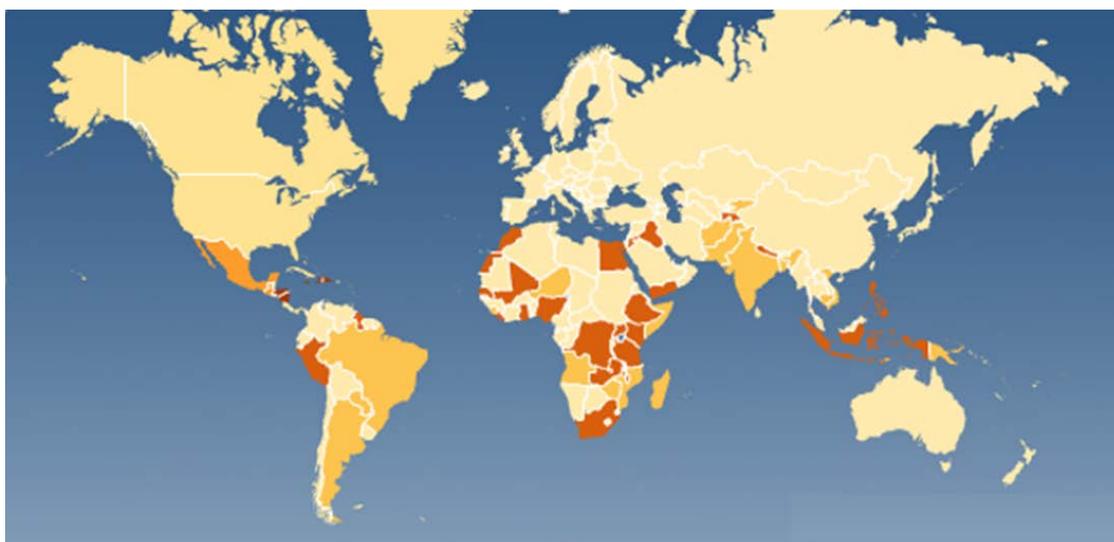
<sup>4</sup> Functional Literacy Assessment Tool developed and used by World Vision:  
<http://www.wvi.org/development/publication/functional-literacy-assessment-tool-flat>

During these early stages of development of the EGRA instrument, a decision was reached to make EGRA open source and readily available to support a higher level and wider dissemination of knowledge on reading and learning outcomes. The purpose behind this decision was to ensure that both technical and nontechnical audiences would become more aware of current education information for their context, and would be able to apply it in making decisions and creating policies.

### 1.3 The Instrument in Action

In 2007, the World Bank supported a pilot of the draft instrument in Senegal (French and Wolof) and The Gambia (English), while USAID supported a pilot in Nicaragua (Spanish). After these initial pilots, use of EGRA expanded across several funders and numerous implementers, countries, and languages. USAID has been one of the largest sponsors of EGRA administrations through the EdData II contract. Between 2006 and mid-2015, EdData II alone supported EGRA studies in 23 countries and 36 languages (**Exhibit 4**).

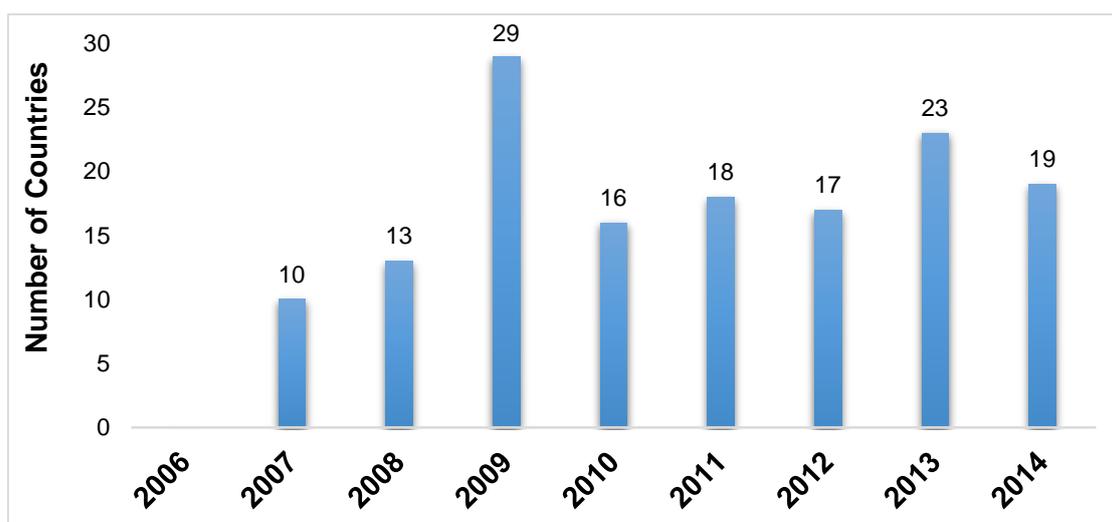
#### Exhibit 4. Map of EGRA administrations



Source: RTI International for the EdData II project website, <https://www.eddataglobal.org/countries/index.cfm>

As of September 2015, nearly 10 years after the birth of EGRA, the tool had been used by over 30 organizations in over 70 countries. The early grade reading approach also shifted to focus on mother-tongue instruction, and as such the instrument has been adapted for administration in over 120 different languages. EdData II has tracked these applications on behalf of USAID; see graph in **Exhibit 5**.

## Exhibit 5. Worldwide application of the EGRA instrument: Number of countries, by year



Data source: RTI International (2015a).

### 1.4 EGRA's Presence in Zambia

USAID/Zambia has supported EGRA data collections on several occasions over the past few years. Starting in 2011, RTI International, with funding and support from USAID, conducted an EGRA in Chibemba in 40 schools across four provinces (Northern, Luapula, Copperbelt, and Central). The purpose of this EGRA application was to provide information to USAID and Zambia's Ministry of General Education (MOGE) about student learning outcomes with regard to literacy in a small sample of schools.

Then, in 2014, RTI International supported the Examinations Council of Zambia (ECZ) to administer the Grade 2 National Assessment Survey (G2 NAS), which included the EGRA. The 2014 survey was administered in 486 schools to a total of 4,855 students. The EGRA was adapted for all seven Zambian national languages. Along with the EGRA, RTI and the ECZ also administered the following tools as part of the G2 NAS: Early Grade Mathematics Assessment (EGMA) survey, funded by the UK Department for International Development [DFID]), student interview questionnaire, teacher interview questionnaire, head teacher interview questionnaire, and classroom and school inventories. The purpose of this survey was to provide information to the MOGE and international donors about student performance in reading and mathematics, as well as school-level factors that impact those outcomes, to contribute to evidence-based decision-making about education policy and practice.

Aside from the abovementioned EGRA applications, USAID/Zambia has also used EGRA before and during implementation of school-level intervention projects to generate impact evaluation data:

- **Read to Succeed (RTS):** In November 2012 and 2014, RTS collected EGRA data from grade 2 and 3 students in 200 government schools in six provinces (Northern, Luapula, Muchinga, Eastern, North Western, and Western) and four

languages (Chibemba, Cinyanja, Kikaonde, and Silozi). Along with the EGRA, RTS also administered the following tools: school data form; head teacher interview and performance checklist; MOGE officials' interview form; teacher interview and performance checklist; and classroom observation form.

- **Time to Learn (TTL):** In November 2012 and 2014, TTL collected EGRA data from 102 community schools in six provinces (Lusaka, Central, Eastern, Copperbelt, Southern, and Muchinga) and three languages (Chinyanja, Chibemba, and Chitonga). EGRA was administered to a maximum of 20 students in grade 2 per school, make the total sample about 1,500 learners. Along with the EGRA, TTL also administered the following tools: community school questionnaire; community school head teacher questionnaire; zonal head questionnaire; grade 2 teacher questionnaire and focus group discussion; standard classroom observation protocol for literacy; learner focus group discussion; and parent community school committee focus group discussion.

## 2 PURPOSE AND USES OF EGRA

### 2.1 History and Overview

Although it was clear from the outset that EGRA would focus on the early grades and the foundational skills of reading, uses of the results were more open to debate.

The original EGRA instrument was primarily designed to be a sample-based “system diagnostic” measure. Its main purpose was to document student performance on early grade reading skills in order to inform governments and donors regarding system needs for improving instruction. Over time, its uses have expanded to include all of the following, with different uses in different contexts:

- Generate baseline data on early reading acquisition in particular grades and/or geographies
- Guide the design of instructional programs by identifying key skills or areas of instruction that need to be improved
- Identify changes in reading levels over time
- Evaluate the outcomes or impact of programs designed to improve early grade reading
- Explore cost-effectiveness of different program designs
- Develop reading indicators and benchmarks
- Serve as a system diagnostic (see Section 2.2) to inform education sector policy, strategic planning, resource allocation

In addition, “the subtasks included in EGRA can be adapted for teachers to inform their instruction<sup>5</sup>. As a formative assessment, teachers can either use EGRA in its entirety or select subtasks to monitor classroom progress, determine trends in performance, and adapt instruction to meet children’s instructional needs” (Dubeck & Gove, 2015, p. 2).

However, to be clear, as it is currently designed, EGRA has its limitations. It is not intended to be a high-stakes accountability measure to determine student grade promotion or to evaluate individual teachers. EGRA is designed to complement, rather than replace, existing curriculum-based pencil-and-paper assessments. EGRA is made up of a set of subtasks that measure foundational skills that have been found to be predictive of later reading success. However, due to the constraints imposed by children’s limited attention span and stamina, neither EGRA nor any other single instrument is capable of measuring all skills required for students to read with comprehension. EGRA is not intended to be an instructional program, but rather is capable of informing instructional programs. EGRA cannot fully determine

---

<sup>5</sup> Using EGRA as a classroom-based formative assessment can be done only with specific required modifications to the instrument and sampling procedures. Classroom-based assessments would also require teachers’ professional development, with specific instructions on administration and interpretation of subtasks.

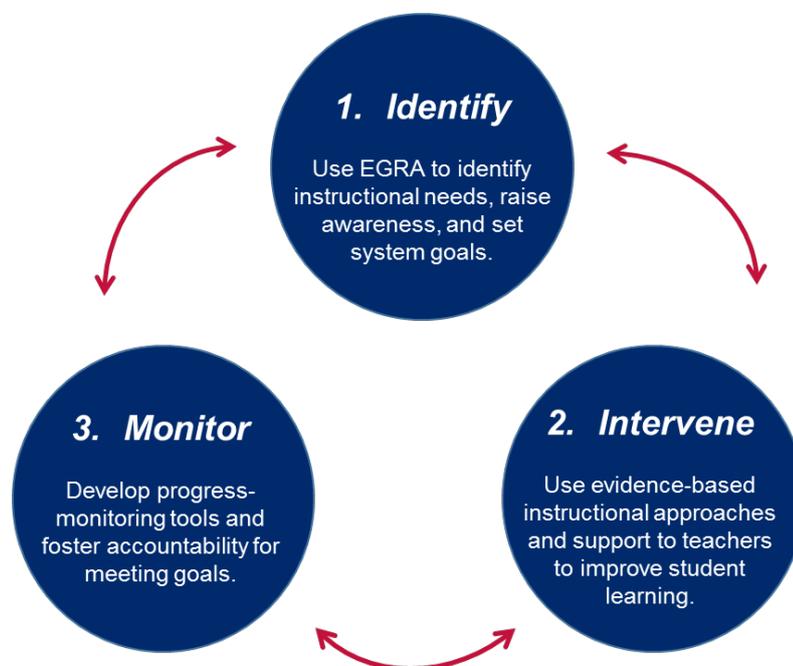
background or literacy behaviors that could impact a student's ability to read (Dubeck & Gove, 2015). Moreover, EGRA's measures are restricted to skills that are subject to influence by instruction, so that the findings will be actionable.

## 2.2 EGRA as a System Diagnostic

The system diagnostic EGRA, as presented in this toolkit, is designed to fit into a complete cycle of learning support and improvement. As depicted in **Exhibit 6**, EGRA can be used as part of a comprehensive approach to improving student reading skills, with the first step being an overall system-level *identification* of areas for improvement. EGRA is able to generate baseline data on early reading acquisition (Gove & Dubeck, 2015). General benchmarking and creation of goals for future applications (see Section 10.1.2) can also be done during the initial EGRA application. Based on EGRA results, education ministries or local systems can then *intervene* to guide the content of new or existing programs using evidence-based instructional approaches to support teachers for improving foundational skills in reading. Results from EGRA can thus inform the design of both pre-service and in-service teacher training programs.

Once recommendations are implemented, parallel forms of EGRA can be used to follow progress and gains in student learning over time through continuous *monitoring*, with the expectation that such a process will encourage teachers and education administrators to ensure students make progress in achieving foundational skills.

### Exhibit 6. The continuous cycle of improving student learning



EGRA and EGRA-based assessments can be used to *identify needs*, *intervene*, and *monitor* progress toward improving student learning outcomes.

When working at the system level, researchers and education administrators frequently begin with student-level data, collected on a sample basis and weighted appropriately, in order to draw conclusions about how the system (or students within the system) is performing. Using average student performance by grade at the system level, administrators can assess where students within the education system are typically having difficulties and can use this information to develop appropriate instructional approaches. Like all assessments whose goal is to diagnose difficulties and improve learning outcomes, in order for a measure to be useful: (1) the assessment relates to existing performance expectations and benchmarks, (2) the assessment correlates with later desired skills, and (3) it must be possible to modify or improve upon the skills through additional instruction (Linan-Thompson & Vaughn, 2007). EGRA meets these requirements as follows.

First, in many high-income countries, teachers (and system administrators) can look to existing national distributions and performance standards for understanding how their students are performing compared to others. In the United States and Europe, by comparing subgroup student performance in relation to national distributions and performance standards, system administrators can decide whether schools and teachers need additional support. In a similar way, EGRA can be used by low-income countries to pinpoint regions (or if the sample permits, schools) that merit additional support, including teacher training or other interventions. When EGRA was first designed, the problem for low-income countries was that similar benchmarks based on locally generated results were not (yet) available. In the meantime, work has been begun in at least 12 countries, including Zambia, to draft national or regional benchmarks using EGRA data. In July 2015, MOGE and the ECZ worked with key partners and experts to use EGRA data from the Grade 2 NAS to define and draft benchmarks for specific skill areas of early grade reading and math. Details are discussed in Section 10.1.

Second, the EGRA tasks were developed intentionally to be predictive of later reading achievement, and numerous administrations of EGRA in multiple countries and languages have generally confirmed the expected correlations. Although the phonological and orthographic variations among languages influence the rate and timing of reading acquisition, all of the skills measured by EGRA have been shown to correlate to reading skills in alphabetic orthographies. As an example, knowing the relationship between sounds and the symbols that represent them has a predictive relationship to success with word reading. Oral reading fluency has been shown to be predictive of reading comprehension. These skills are measured in EGRA and, therefore, we can assume with confidence that EGRA results relate something meaningful about the direction in which the children are headed in the reading acquisition process.

Third, EGRA not only can give us meaningful predictions about future performance, but also can direct our attention to needed instructional changes. It makes little sense to measure something that we have no hope of changing through additional instruction. EGRA is valuable as a diagnostic tool precisely because it includes measures of those skills that can be improved through instruction.

# 3 CONCEPTUAL FRAMEWORK AND RESEARCH FOUNDATIONS

The conceptual framework of reading acquisition underpinning the development of EGRA is guided by the work of the U.S. National Reading Panel (National Institute of Child Health and Human Development, 2000), August and Shanahan (2006), and the Committee on the Prevention of Reading Difficulties in Young Children (Snow, Burns, & Griffin, 1998), among others. The extensive literature on reading points to the need for students to acquire specific skills through targeted instruction in order to become successful lifelong readers.

## 3.1 Summary of Skills Necessary for Successful Reading

The ultimate goal of learning to read is comprehension, or “the process of simultaneously extracting and constructing meaning through interaction and involvement with written language” (Snow & the RAND Reading Study Group, 2002, p. 11). To competent readers, reading may seem effortless; they read a text and understand it with such speed and ease that they are not conscious of the process of comprehension itself. However, comprehension is actually a highly complex skill that is built from a wide array of subskills working together simultaneously.

Reading acquisition is seen as a developmental process (Chall, 1996). Higher-order skills (e.g., fluency and comprehension) build on lower-order skills (e.g., phonemic awareness, letter sound knowledge, and decoding), and the lower-order skills have been shown to be predictive of later reading achievement. Therefore, even if children cannot yet read a passage with comprehension, we can nonetheless measure their progress toward acquiring the lower-order skills that are necessary steps along the path to that end.

Five components are generally accepted as necessary to master the process of reading: phonological awareness, phonics (method of instruction that helps teach sound–symbol relationships), vocabulary, fluency, and comprehension (Armbruster, Lehr, & Osborn, 2003; Vaughn & Linan-Thompson, 2004). The skills within each component are not sufficient on their own to produce successful reading, but they build on one another and work together to reach the ultimate goal of reading—i.e., comprehension. The EGRA subtasks (refer to Section 4) are aligned to these components of reading. Because these skills are acquired in phases, at any given point in time, some subtasks are likely to have floor effects (that is, most children in the early grades would not be able to perform at a sufficient skill level to allow for analysis) and others ceiling effects (almost all children receive high scores), depending on where the children are in their development.

## 3.2 Phonological Awareness

### 3.2.1 Description

Phonological awareness can be defined as “the ability to detect, manipulate, or analyze the auditory aspects of spoken language (including the ability to distinguish or segment words, syllables, or phonemes), independent of meaning” (National Center for Family Literacy [NCFL], 2008, p. vii). *Phonemic awareness*, a term often used interchangeably with *phonological awareness*, is actually a subset thereof and refers specifically to the awareness of *phonemes*, which are the smallest units of sound that distinguish the meaning of a word in a given language. For example, the English consonant sounds /p/<sup>6</sup>, /k/, and fricative /ð/ (i.e., the “th” sound) are the phonemes that make the word “pat” distinguishable from “cat” and “that” in spoken language.

Similarly, in alphabetic orthographies, a *grapheme* is to written language what a phoneme is to oral language—that is, as explained in the glossary at the beginning of the toolkit, it is “the most basic unit in an alphabetic written system that can change the meaning of a word. A grapheme might be composed of one or more than one letter; or of a letter with a diacritic mark.” Languages vary in the degree of direct correspondence between phonemes and graphemes; in some languages, like Spanish, graphemes and phonemes have nearly a one-to-one correspondence, but in English, the mapping is much more complex. For example, in English the phoneme /k/ may be spelled with the letters *c*, *k*, *ck*, *ch*, *qu*, etc., just as the letter *c* may represent the phoneme /k/ in one word and /s/ in another.

As humans process rapid oral language input, our phonological knowledge remains, for the most part, efficiently subconscious. Learning to read (in alphabetic orthographies), however, requires linking graphemes to individual phonemes, which requires a conscious awareness of the phonemes in the language and the ability to distinguish between and manipulate them (Gove & Wetterberg, 2011). Phonological awareness enables children to separate words into sounds and blend sounds into words, oral skills that are necessary precursors to decoding and spelling.

Research suggests that children’s awareness of speech sounds develops progressively, beginning with larger units—i.e., at the word level—then moving to the smaller units of the syllable, onset–rime (beginning and ending sounds), and finally, the phoneme. In fact, sensitivity to the phoneme level, which is essential for word decoding, may not begin to develop until the onset of literacy instruction (Goswami, 2008). Phonological awareness has been shown across numerous studies in multiple languages to be predictive of later reading achievement (Badian, 2001; Denton, Hasbrouck, Weaver, & Riccio, 2000; Goikoetxea, 2005; McBride-Chang & Kail, 2002; Muter, Holme, Snowling, & Stevenson, 2004; Wang, Park, & Lee, 2006).

---

<sup>6</sup> Phonemes are traditionally written between slashes in the International Phonetic Alphabet. The full IPA chart is available for reference and use from <http://www.internationalphoneticassociation.org/content/ipa-chart>, under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2005 International Phonetic Association.

### 3.3 The Alphabetic Principle, Phonics, and Decoding

#### 3.3.1 Description

The alphabetic principle is the understanding that words are made up of sounds (i.e., phonemes) and that letters (i.e. graphemes) are symbols that represent those sounds. The Alphabetic Principle is an abstract concept which is best taught explicitly to students in order to clarify what the symbols on the page represent in their most elemental forms. When students understand that sounds map onto letters, they can begin to learn to decode words. Alphabet knowledge includes knowledge of the individual letter names, their distinctive graphic features, and which phoneme(s) each represents.

Teaching these grapheme-to-phoneme and phoneme-to-grapheme mappings is an instructional method commonly known as phonics. Research has shown alphabet knowledge to be a strong early predictor of later reading achievement (Adams, 1990; Ehri & Wilce, 1985; Piper & Korda, 2010; Wagner, Torgesen, & Rashotte, 1994; Yesil-Dağlı, 2011), for both native and nonnative speakers of a language (Chiappe, Siegel, & Wade-Woolley, 2002; McBride-Chang & Ho, 2005; Manis, Lindsey, & Bailey, 2004; Marsick & Watkins, 2001). One of the main differences between successful readers and struggling readers is their ability to use the letter–sound correspondences to decode new words they encounter in text and to encode (spell) the words they write (Juel, 1991).

## LANGUAGE PHONOLOGIES AND ORTHOGRAPHIES

Languages vary in the complexities of their **phonologies** (sound systems); some languages have many more phonemes than others, some allow much more complex syllable structures (e.g. with consonant clusters in initial and final position), some have much longer words on average than others, etc. Likewise, **orthographies** (spelling system of a language) vary in the degree of transparency or consistency of the letter-sound relationships.

In highly transparent orthographies, the correspondence between phonemes and graphemes is nearly one-to-one. This facilitates their acquisition because almost every letter will reliably represent one and the same sound regardless of the word in which it appears, and vice versa. By contrast, English has what is called an “opaque” or “deep” orthography, because nearly every letter maps to more than one sound and every sound to more than one letter, thereby complicating the mapping process considerably.

In brief, both the relative complexity of the phonological system of a given language and its orthography have consequences for the rate of acquisition of related reading subskills such as phonics. At the two extremes, a child learning to read in a consistent, transparent orthography of a language with relatively low phoneme inventory, simple syllable structures, and short average word lengths will be at an advantage for mastering the letter–sound mappings and decoding skills more rapidly than a child learning to read in a language with an opaque orthography, many irregularities, many phonemes, complex syllable structures, and long average word lengths. This is one reason why cross-linguistic benchmarks as well as comparisons of EGRA findings are not appropriate.

According to the “dual route” model of word recognition (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Zorzi, 2010), there are two distinct but not mutually exclusive ways in which humans process text to recognize words. They are referred to as the lexical and sublexical routes.

Reading via the lexical route involves looking up a word in the mental lexicon containing knowledge about spellings and pronunciations of real words. “Instant word recognition” means that the word on the page is familiar and instantly recognizable because of knowledge of the letter strings and spelling pattern. In the sublexical route, we decode the word by converting the letters into sounds using our knowledge of their mappings, blend the sounds into a word, and then recognize the word based on its phonological form.

The lexical route may be faster for familiar words, and is necessary for processing words with irregular spellings, but the sublexical route is necessary for processing new or unfamiliar words. In languages with highly consistent orthographies (and therefore few irregular spellings), all words are essentially decodable and accessible through the sublexical route. EGRA uses the nonword reading task to assess student skills in decoding via the sublexical route.

### 3.3.2 Measures of Alphabet Knowledge and Decoding Skills

EGRA assesses children’s alphabet knowledge in several ways, beginning with the **letter sound identification** subtask, a component of the core EGRA. The letter sound identification subtask tests children’s ability to recognize the graphemic features of each letter and accurately map it to its corresponding name or sound. Children are given a written list of capital and lowercase letters (and diphthongs or digraphs if appropriate) in random order and asked to articulate either the name or the sound of each.

The next step up in skill difficulty is for readers to use their mastery of the letter–sound correspondences to decode words. Therefore, the **nonword reading** subtask, another core EGRA subtask, provides indirect insight into children’s ability to decode unfamiliar words. The nonword reading subtask presents the children with a written list of pseudowords that follow the phonological and spelling rules of the language but are not actual words in the language. Children are asked to read out loud as many of the nonwords as they can, as quickly and carefully as they can. According to the dual-route model, this subtask requires children to apply their decoding skills based on their knowledge of the grapheme-phoneme mappings. Because nonwords will not have any whole-word representation previously stored in long-term memory to be accessed directly, students must rely on decoding in order to identify them.

## 3.4 Vocabulary and Oral Language

### 3.4.1 Description

Reading comprehension involves more than just word recognition. In order to construct meaning, we must link the words we read to their semantic representation or meaning attached to the word in our minds; and knowing the meaning of words

relates to one's overall oral language comprehension (Kamhi & Catts, 1991; Nation, 2005; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). Vocabulary refers to the ability to understand the meaning of words when we hear or read them (receptive), as well as to use them when we speak or write (productive). Reading experts have suggested that vocabulary knowledge of between 90 and 95 percent of the words in a text is required for comprehension (Nagy & Scott, 2000). It is not surprising, then, that in longitudinal studies, vocabulary has repeatedly been shown to influence and be predictive of later reading comprehension (Muter et al., 2004; Roth, Speece, & Cooper, 2002; Share & Leiken, 2004).

### 3.4.2 Measures of Vocabulary

Although none of the core EGRA subtasks measures vocabulary directly, an optional, untimed **vocabulary** subtask measures receptive-language skills of individual words and phrases related to body parts, common objects, and spatial relationships. This subtask has been used in a few contexts but has not yet been through the same expert panel review and validation process as the other subtasks.

In addition, **listening comprehension**, which is a core EGRA subtask, assesses overall oral language comprehension, and therefore, indirectly, oral vocabulary on which it is built in part. For this subtask, assessors read children a short story on a familiar topic and then ask children three to five comprehension questions about what they heard. The listening comprehension subtask is used primarily in juxtaposition with the reading comprehension subtask (see Comprehension, Section 3.6 below) in order to tease out whether comprehension difficulties stem primarily from low reading skills or from low overall language comprehension.

## 3.5 Fluency

### 3.5.1 Description

Fluency is “the ability to read text quickly, accurately, and with proper expression” (NICHD, 2000, pp. 3–5). According to Snow and the RAND Reading Study Group (2002):

Fluency can be conceptualized as both an antecedent to and a consequence of comprehension. Some aspects of fluent, expressive reading may depend on a thorough understanding of a text. However, some components of fluency—quick and efficient recognition of words and at least some aspects of syntactic parsing [sentence structure processing]—appear to be prerequisites for comprehension. (p. 13)

Fluency can be seen as a bridge between word recognition and text comprehension. While decoding is the first step to word recognition, readers must eventually advance in their decoding ability to the point where it becomes automatic; then their attention is free to shift from the individual letters and words to the ideas themselves contained in the text (Armbruster et al., 2003; Hudson, Lane, & Pullen, 2005; LaBerge & Samuels, 1974). Speed may also be critical due to the constraints of our short-term

working memory. Working memory can only hold so much information at one time, and if we decode too slowly because we are paying attention to each individual word part, we will not have enough space in our working memory for the whole sentence; we will forget the beginning of the text sequence by the time we reach the end. If we cannot hold the whole sequence in our working memory at once, we cannot extract meaning from it (Abadzi, 2006; Hirsch, 2003).

Like comprehension, fluency itself is a higher-order skill requiring the complex and orchestrated processes of decoding, identifying word meaning, processing sentence structure and grammar, and making inferences, all in rapid succession (Hasbrouck & Tindal, 2006). It develops slowly over time and only from considerable exposure to connected text and decoding practice.

Numerous studies have found that reading comprehension correlates to fluency, especially in the early stages (Fuchs, Fuchs, Hosp, & Jenkins, 2001) and for individuals learning to read in a language they speak and understand. For example, tests of oral reading fluency, as measured by timed assessments of correct words per minute, have been shown to have a strong correlation (0.91) with the reading comprehension subtest of the Stanford Achievement Test (Fuchs et al., 2001). Data from many EGRA administrations across contexts and languages have confirmed the strong relationship between these two constructs (Bulat et al., 2014; LaTowsky, Cumiskey, & Collins, 2013; Management Systems International, 2014; Pouezevara, Costello, & Banda, 2012; among many others). The importance of fluency as a predictive measure does, however, decline in the later stages as students learn to read with fluency and proficiency. As students become more proficient and automatic readers, vocabulary becomes a more important predictor of later academic success (Yovanoff, Duesbery, Alonzo, & Tindall, 2005).

How fast is fast enough? While it is theorized that a minimum degree of fluency is needed in order for readers to comprehend connected text, fluency benchmarks will vary by grade level and by language. A language with shorter words on average, like English or Spanish, allows students to read more words per minute than a language like Kiswahili, where words can consist of 10–15 or even 20 letters. In other words, the longer the words and the more meaning they relay, the fewer the words that need to be read per minute.

### 3.5.2 Measures of Fluency

Given the importance of fluency for comprehension, EGRA's most direct measurement of fluency, the **oral reading fluency with comprehension** subtask, is a core component of the instrument. Children are given a short written passage on a familiar topic and asked to read it out loud "quickly but carefully." Fluency comprises speed, accuracy, and expression (prosody). The oral reading fluency subtask is timed and measures speed and accuracy in terms of the number of correct words read per minute. This subtask does not typically measure expression.

Besides the oral reading fluency subtask, several other EGRA subtasks discussed above are timed and scored for speed and accuracy in terms of correct letters (or sounds and syllables) or words per minute: letter name identification, letter sound identification, nonword reading, and familiar word reading. Because readers become increasingly more fluent as their reading skills develop, timed assessments help to

track this progress across all these measures and show where children are on the path to skilled reading.

## 3.6 Comprehension

### 3.6.1 Description

Comprehension is the ultimate goal of reading. It enables students to make meaning out of what they read and use that meaning not only for the pleasure of reading but also to learn new things, especially other academic content. Reading comprehension is also a highly complex task that requires both extracting and constructing meaning from text. Reading comprehension relies on a successful interplay of motivation, attention, strategies, memory, background topic knowledge, linguistic knowledge, vocabulary, decoding, fluency, and more, and is therefore a difficult construct for any assessment to measure directly (Snow & the RAND Reading Study Group, 2002).

### 3.6.2 Measures of Reading Comprehension

EGRA measures reading comprehension through the **oral reading passage** subtask, based on the short paragraph that children read aloud for the oral reading fluency subtask. After children read the passage aloud, they are asked three to five comprehension questions, both explicit and inferential, that can be answered only by having read the passage. Lookbacks—i.e., referencing the passage for the answer—may be permitted to reduce the memory load but are not typically used in the core instrument.

# 4 EGRA INSTRUMENT DESIGN: ADAPTATION DEVELOPMENT AND ADAPTATION MODIFICATION

This section discusses the structure and requirements necessary for designing or modifying an EGRA for any given context. The text throughout this section of the toolkit exposes readers to the various subtasks that can be included in an EGRA instrument by providing subtask descriptions and specific construction guidelines.

## 4.1 Adaptation Workshop

The first adaptation step is to organize an in-country workshop, normally lasting about five working days. This subsection reviews the steps for preparing and delivering an EGRA adaptation workshop and provides an overview of the topics to be covered.

This in-country adaptation workshop is held at the start of the test development (or modification) process for EGRA instruments. It provides an opportunity for countries to build **content validity** (see glossary) into the instrument by having government officials, curriculum experts, and other relevant groups examine the EGRA subtasks and make judgments about the appropriateness of each item type for measuring the early reading skills of their students, as specified in curriculum statements or other guidelines that state learning expectations or standards.<sup>7</sup> As part of the adaptation process, the individuals participating in the workshop adapt the EGRA template as necessary and prepare country-appropriate items for each subtask of the test. This approach ensures that the assessment has **face validity** (see glossary). Following the workshop, piloting of the instrument in a school (in teams) is essential. Pilot testing and fieldwork are discussed in detail in Section 7.

For additional information on the technical quality and reliability of the EGRA instrument, including guidelines for conducting basic instrument quality and reliability checks, please see Section 7.1.2 of this toolkit.

The objectives of the workshop are:

- Give both government officials and local curriculum and assessment specialists a grounding in the research backing of the instrument components.
- Adapt the instrument to local conditions using the item-construction guidelines provided in this toolkit, including

---

<sup>7</sup> The degree to which the items on the EGRA test are representative of the construct being measured is known as **test-content-related evidence** (i.e., early reading skills in a particular country).

- translating the instrument instructions;
- developing versions in appropriate languages, if necessary; and
- modifying the word and passage reading components to reflect locally and culturally appropriate words and concepts.

**Exhibit 7** more clearly defines the differences between development and modification workshops. If a country-specific EGRA is being developed for the first time, it is considered an **adaptation development**; if EGRA has already been conducted in country, then the workshop is an **adaptation modification**.

### Exhibit 7. Differences between EGRA adaptation development and adaptation modification

Adaptation (development) of new instruments	Adaptation (modification) of existing instruments
Language analysis	Language analysis (optional)
Item selection	Item reordering/randomization
Verification of instructions	Verification of instructions
Pretesting	Pretesting

#### 4.1.1 Overview of Workshop Planning Considerations

Whether designing a country-specific EGRA instrument from the beginning (development) or from an existing model (modification), the team will need to make sure the instrument is appropriate for the language(s), the grade levels involved in the study, and the research questions at hand.

The development of the instrument will require a selection of appropriate subtasks and subtask items. Further considerations include:

- The agenda must allow for limited field testing of the instrument as it is being developed, which includes taking participants (either a subgroup or all) to nearby schools to use the draft instrument with students. This field testing allows participants a deeper understanding of the instrument as well a rough test of the items to gauge any obvious changes that may be needed (such as revisions to ambiguous answer choices or overly difficult vocabulary).
- Some of the language analysis that is necessary to draft items can be done in advance, along with translation of the directions, which must remain standardized across countries. For purposes of standardization, all students must be given the same opportunities regardless of assessor or context; therefore, it is required to keep the instructions the same across all countries and contexts.
- If the workshop cannot be done in the region where testing will take place, the study team must arrange for a field test afterward, or find a group of nearby students who speak the language and who are willing to participate in a field test during the workshop. For either arrangement, the field test team will need to monitor the results and report back to the full group.
- The most difficult part of adaptation is usually story writing, so it is important not to leave this subtask until the last day. This step involves asking local experts to

write short stories using grade-level appropriate words, as well as to write appropriate comprehension questions to accompany the stories. Both the stories and the questions often need to be translated into English or another language for review by additional early grade reading experts, and revised multiple times in the language of assessment before finalization.

#### 4.1.2 Who Participates?

Groups composed of government staff, teacher trainers, former or current teachers, and language experts from local universities offer a good mix of experience and knowledge—important elements of the adaptation process. However, the number of participants in the adaptation workshop will be determined by the availability of government staff to participate. Their presence is recommended in order to build capacity and help ensure sustainability for the assessment. The number of participants will depend in part on the number of languages involved in the adaptation process for a given study, but in general, 30 is a recommended maximum.

Workshop participants always include:

1. Language experts: To verify the instructions that have been translated, to guide the review of items selected, and to support the story writing or modifications
2. Nongovernment practitioners: Academics (reading specialists, in particular), and current or former teachers (with a preference for reading teachers)
3. Government officials: Experts in curriculum development, assessment
4. A psychometrician or test-development experts

Ideally, key government staff participate throughout the entire adaptation, assessor training, and piloting process (spread over one month in total, depending on the number of schools to be sampled). Consistency among participants is needed so the work goes forward with clarity and integrity while capacity and sustainability are built.

The workshop is facilitated by a team of at least two experts. Both workshop leaders must be well versed in the components and justifications of the assessment and be adept at working in a variety of countries and contexts.

- **Assessment expert**—is responsible for leading the adaptation (be it development or modification) of the instrument and later, guiding the assessor training and data collection; has a background in education survey research and in the design of assessments/tests. This experience includes basic statistics and a working knowledge of spreadsheet software such as Excel and a statistical program such as SPSS or Stata.
- **Early literacy expert**—is responsible for presenting reading research and pedagogical/instruction processes; has a background in reading assessment tools and instruction.

#### 4.1.3 What Materials Are Needed?

Materials for the adaptation workshop include:

- Paper and pencils with erasers for participants

- LCD projector, whiteboard, and flipchart (if possible, the LCD projector should be able to project onto the whiteboard for simulated scoring exercises)
- Current national or local reading texts, appropriate for the grade levels and the languages to be assessed (these texts will inform the vocabulary used in story writing and establish the level of difficulty)
- Paper copies of presentations and draft instruments
- Presentation on the EGRA-related reading research, development process, purpose, uses, and research background
- Samples of EGRA oral reading fluency passages, comprehension questions, and listening comprehension questions from other countries; or for modification, copies of the previous in-country EGRA instrument.

A sample agenda for the adaptation and research workshop is presented in **Exhibit 8**.

### Exhibit 8. Sample agenda: EGRA adaptation development or adaptation modification workshop

Day & Time	Day 1	Day 2	Day 3	Day 4	Day 5
9:00-9:30 a.m.	Welcome and introduction	Review of Day 1	Review of Day 2	Review of Day 3	Visit schools to field test instruments and questionnaires
9:30-10:30 a.m.	Project overview and EGRA context	Review draft EGRA instrument (e.g., non-words)	Development of Listening Comprehension Passages	Modify/develop additional subtasks and questionnaires, as applicable	
10:30-11:00 a.m.	<i>Break</i>				
11:00-12:30 p.m.	Overview of EGRA: purpose, instrument content, results use	Development of Oral Reading Fluency Passages	Continue listening comprehension stories and develop questions	Modify/develop additional subtasks and questionnaires, as applicable	School visit debrief
12:30-1:30 p.m.	<i>Lunch</i>				
1:30-3:00 p.m.	Presentation on language: orthography and issues to consider vis-à-vis EGRA development	Continue ORF stories and develop questions	Review and Update Pupil Questionnaire	Review and practice EGRA administration for field test	Finalization of instruments
3:00-3:45 p.m.	<i>Break</i>				
3:45-5:00 p.m.	Review draft EGRA instrument: (e.g., phonemic awareness and letter sounds)	Finalize stories and questions	Finalize stories, questions, pupil questionnaire as needed	Review and practice EGRA administration for field test	Workshop Closure
Daily Objectives:	<i>Understanding of EGRA purpose and content</i>	<i>Oral reading passages and questions developed</i>	<i>Listening comprehension passages and stories developed; Pupil Questionnaire Developed</i>	<i>Additional subtasks/questionnaires developed</i>	<i>Instruments finalized</i>

NOTE: The duration of the adaptation workshop and specific sessions will depend on several factors, including: existence of a previously used EGRA for the given language/country/grade; number of subtasks to be tested; number of languages to be tested; need for additional questionnaires and instruments; and purpose and audience of the workshop.

## 4.2 Review of the Zambian Instrument Components

As discussed in Section 1, the initial EGRA design was developed with the support of experts from USAID, the World Bank, and RTI. Over the years, expert consultations have led to a complete Early Grade Reading Assessment application in English that has been continually reviewed and updated. The instrument being used in Zambia is an adapted version of the standard EGRA, which combines the following subtasks (some of which are administered in a local language and others which are administered in English):

1. Listening comprehension (administered in local language and English)
2. Letter Sound identification
3. Nonword reading
4. Oral reading passage with reading comprehension
5. Orientation to print
6. Vocabulary

It is important to note that the instrument and procedures presented here have been demonstrated to be a reasonable starting point for assessing early grade reading (see NICHHD, 2000; and Dubeck & Gove, 2015). That is, the skills measured by the EGRA are essential but not sufficient for successful reading: EGRA covers a significant number of the predictive skills but not all skills or variables that contribute to reading achievement. For example, EGRA does not measure a child's background knowledge, motivation, attention, memory, reading strategies, productive vocabulary, comprehension of multiple text genres, retell fluency, etc. No assessment can cover all possible skills, as it would be exceptionally long, causing students to become fatigued and perform poorly. The instrument should not be viewed as sacred in terms of its component parts, but it is recommended that variations, whether in the task components or in the procedures, be justified, documented in terms of the purpose and use of the assessment, and shared with the larger community of practice.

### Exhibit 9. Review of Zambian instrument components

Subtask	Early reading skill	Skill demonstrated by students' ability to:
1. Listening comprehension <sup>8</sup>	Listening comprehension; oral language	Respond correctly to different types of questions, including literal and inferential questions about the text the assessor reads to them
2. Letter identification: letter sounds	Alphabet knowledge	Provide the sound of letters presented in both upper case and lower case in a random order
3. Nonword reading	Decoding	Make letter–sound (grapheme-phoneme correspondences, or GPCs) through the reading of simple nonsense words
4. Oral reading passage with reading comprehension	Oral reading fluency Reading comprehension	Read a text with accuracy, with little effort, and at a sufficient rate Respond correctly to different types of questions, including literal and inferential questions about the text they have read
5. Orientation to print	Concepts about print; print awareness	Indicate text direction or other basic knowledge of print
6. English Vocabulary	Vocabulary Knowledge	

#### 4.2.1 Listening Comprehension

A listening comprehension assessment involves a passage that is read aloud by the assessor, and then students respond to oral comprehension questions or statements.

<sup>8</sup> The Zambian EGRA instrument includes two listening comprehension subtasks: one story and corresponding questions are read aloud in the local language; a second (different) story and corresponding questions are administered in English.

This subtask can be included at the beginning of the series to ease the children into the assessment process and orient them to the language of assessment.

Testing listening comprehension separately from reading comprehension is important because it provides information about what students are able to comprehend without the challenge of decoding a text. Students who are struggling or have not yet learned to decode may still have oral language, vocabulary, and comprehension skills and strategies that they can demonstrate apart from reading text. This gives a much fuller picture of what students are capable of when it comes to comprehension. Listening comprehension tests have been around for some time and in particular, have been used as an alternative assessment for disadvantaged children with relatively reduced access to print (Orr & Graham, 1968). Poor performance on a listening comprehension tool suggests either that children lack basic knowledge of the language in question, or that they have difficulty processing what they hear.

**Data.** Students are scored on the number of correct answers they give to the questions asked (out of the total number of questions). Instrument designers avoid questions with only “yes” or “no” answers.

**Item construction.** Passage length depends on the level and first language of the children being assessed, although most passages are approximately 30 words in length in order to provide enough text to develop material for three to five comprehension questions. The story narrates a locally adapted activity or event that will be familiar to the children. The questions must be similar to the questions asked in the reading comprehension task (described below). Most will be literal questions that can be answered directly from the text. One or two questions are inferential, requiring students to use their own knowledge as well as the text to answer the question.

**Exhibit 10** and **Exhibit 11** are samples of the listening comprehension subtask.

## Exhibit 10. Sample: Listening comprehension (English)

EGRA (English) Assessment (LANGUAGE)

Sub-test 5. LISTENING COMPREHENSION		📖 X		🕒 X
<p>🔊 <b>Merebekenkan abasem tiawa bi baako pe dennennen akyere wo, na mabisabisa wo nsem kakra afa ho. Mesre wo tie no yiye na bua nsemmisa no senea wubetumi biara. Wubetumi de kasa biara a wope ayiyi nsemmisa no ano. Metumi afa ase? Yemfi ase.</b> I am going to read you a short story aloud ONCE and then ask you some questions. Please listen carefully and answer the questions as best as you can. You can answer the questions in whichever language you prefer. Ready? Let's begin.</p>			<p>Remove the pupil stimuli booklet from the child's view.</p> <p>Do not allow the child to look at the passage or the questions.</p> <p>If a child says "I don't know," mark as incorrect.</p>	
<p>👁️ (✓) 1 = Correct (✓) 0 = Incorrect (✓) . = No response.</p>				
<p><b>Issa was very sad. He lost his grandfather's sheep. He could not go to look for them. Grandfather came to look for them. Soon he returned with the sheep. Issa is smiling now.</b></p>				
<p><b>Why was Issa sad?</b> [he lost his sheep; he could not go to look for his sheep]</p>	1	0	.	
<p><b>Who went to look for the sheep?</b> [Grandfather]</p>	1	0	.	
<p><b>Why is Issa smiling now?</b> [Grandfather returned with his sheep; his sheep are back; Grandfather found the sheep]</p>	1	0	.	

**Mo! Woaye ade. Yenko ofa a edi so no so.** Good effort! Let's go on to the next section.

## Exhibit 11. Sample: Listening comprehension (Chinyanja)

Sub-test 1. LISTENING COMPREHENSION		X		X	
<p><b>🔊</b> Ndidzakwerengera ka nthano/nkhani mokweza KAMODZI ndipo pambuyo pake ndidzakufunsa mafunso. Conde umvetserere mosamalitsa ndipo uyankhe mafunso mmene ungakwanitsire. Ungayankhe mafunso mcilankhulo ciriconse cimene ukonda. I am going to read you a short story aloud ONCE and then ask you some questions. Please listen carefully and answer the questions as best as you can. You can answer the questions in whichever language you prefer. Ready? Let's begin.</p>				<p>Remove the pupil stimuli booklet from the child's view.</p>	
<p>🔍 (✓) 1 = Correct (✓) 0 = Incorrect (✓) . = No response.</p>				<p>Do not allow the child to look at the passage or the questions.</p>	
<p>Patsiku Lolemba, Mangani anapita kusukulu. Ananyamula mabuku ndi nyama m'cola cake. Pamene anali kuyenda, anapeza galu wamkulu panjira. Anafuna kuthawira pathengo koma anagwa pansi. Yunifomu yake inada ndipo galu anatenga nyama yake. Mangani anathawira kunyumba. Pamene anafika kunyumba, m'bale wake anamubwereka yunifomu yake. Anakondwera.</p>				<p>If a child says "I don't know," mark as incorrect.</p>	
<p><b>Ndi tsiku liti pamene Mangani anapita kusukulu?</b> (Pa Lolemba)</p>		1	0	.	
<p><b>Ananyamula ciani mu cola cake?</b> (Mabuku ndi nyama)</p>		1	0	.	
<p><b>N'ciani cimene anapeza panjira?</b> (Anapeza galu wamkulu)</p>		1	0	.	
<p><b>Ndi cifukwa ciani Mangani anathawa galu?</b> (Anaopa kuti galu angamulume)</p>		1	0	.	
<p><b>Ndi cifukwa ciani m'bale wake anamubwereka yunifomu Mangani?</b> (Cifukwa yunifomu yake inada).</p>		1	0	.	
<p><b>Wacita bwino! Tiye tipitirize patsamba lotsatira</b> Good effort! Let's go on to the next section.</p>					

### 4.2.2 Letter Sound Identification

Knowledge of how letters correspond to sounds is another critical skill children must master to become successful readers. Letter–sound correspondences are typically taught through phonics-based approaches. Letter-sound identification tests the actual knowledge students need to have to be able to decode words—i.e., knowing the sound the letter represents allows students to sound out a word.

In this subtask, students are asked to produce the sounds of all the letters, plus digraphs and diphthongs (e.g., in English: th, sh, ey, ea, ai, ow, oy), from the given list, within a one-minute period.

For letters, the full set of letters of the alphabet is listed in random order, 10 letters to a row, using a clear, large, and familiar font. For example, Century Gothic in Microsoft Word is similar to the type used in many children's textbooks; also SIL International has designed a font called Andika specifically to accommodate

beginning readers.<sup>9</sup> The number of times a letter is repeated is based on the frequency with which the letter occurs in the language in question. The complete alphabet (using a proportionate mixture of both upper and lower case) is presented based on evidence from European languages that student reading skills advanced only after about 80 percent of the alphabet is known (Seymour, Aro, & Erskine, 2003).

Letter-frequency tables will depend on the text being analyzed (a report on x-rays or xylophones will necessarily show a higher frequency of the letter x than the average text). Test developers constructing instruments in other languages sample 20–30 pages of a grade-appropriate textbook or supplementary reading material and analyze the frequency of letters electronically to develop similar letter frequency tables.

Developing a letter-frequency table requires typing the sampled pages into a word-processing program and using the “Find” command. Enter the letter “a” in the “Find what” search box and set up the search to highlight all items found in the document. In the case of Microsoft Word, it will highlight each time the letter “a” appears in the document and will report the number of times it appeared (in the case of this section of the toolkit, for example, the letter “a” appears over 3,500 times). The analyst will repeat this process for each letter of the alphabet, recording the total number for each letter until the proportion of appearances for each letter can be calculated as a share of the total number of letters in the document.

Pronunciation issues need to be handled with sensitivity in this and other subtasks. The issue is not to test for “correct” pronunciation. The assessment tests automaticity using a pronunciation that may be common in a given region or form of the language of the adaptation. Thus, regional accents are acceptable in judging whether a letter sound is pronounced correctly.

For letters that can represent more than one sound, several answers will be acceptable. During training, assessors and supervisors, with the help of language experts, carefully review possible pronunciations of each letter and come to agreement on acceptable responses, giving careful consideration to regional accents and differences. (For a complete listing of characters and symbols in international phonetic alphabets, please see the copyrighted chart created and maintained by the International Phonetic Association at <http://westonruter.github.io/ipa-chart/keyboard/>.)

**Data.** The child’s score for this subtask is calculated as the number of correct letter sounds read per minute. If the child completes all of the letter sounds and digraphs/diphthongs before the time expires, the time of completion is recorded and the calculations based on that time period. In the event that paper assessments must be used, assessors mark any incorrect letters with a slash (/), place a bracket (]) after the last letter named, and record the time remaining on a stopwatch at the completion of the exercise. Electronic data capture does the marking and calculations automatically based on assessors’ taps on the tablet screen. Three data points are used to calculate the total correct letter sounds and diphthongs/digraphs per minute (clspm):

---

<sup>9</sup> More about Andika, including how to download this font, can be found on SIL’s website: [http://scripts.sil.org/cms/scripts/page.php?site\\_id=nrsi&id=andika](http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=andika)

$$\text{clspm} = (\text{Total letter sounds identified} - \text{Total incorrect}) / [(60 - \text{Time remaining on device}) / 60]$$

Each of these data points can also be used for additional analyses. For example, information on the total number of sounds identified will allow for differentiation between a student who names 50 sounds within a minute but names only half of them correctly; and a student who names only 25 sounds within a minute, but names all of them correctly.

Note that this subtask, as well as many of the subtasks that follow it, is not only timed but also time-limited (i.e., stopped after a specified period, whether completed or not). The time limitation is useful in making the assessment shorter, and is also less stressful for both child and assessor, as the child does not have to keep trying to do the whole task at a slow pace. In addition, timing helps to assess automaticity.

**Item construction.** This subtask consists of 100 total items. Letters of the alphabet are distributed randomly, with 10 letters to a line in horizontal rows, and evenly distributed among upper- and lowercase letters. Most of the characters will be presented multiple times. The percentages calculated in the exercise above act as a guide for the frequency with which the letters, diphthongs, and/or digraphs appear in the task sheet.

It is not uncommon for an existing EGRA instrument to need to be modified into one or more parallel versions, for example, for purposes of monitoring gains from baseline to midterm or endline. Under such scenarios, items in some subtasks are reordered, or re-randomized, to create new grids—e.g., 10 rows of 10 letters—without frequencies having to be recalculated. In these cases, to ensure equivalent test forms, it is important that the reordering occur only within the individual rows (in order to retain relative subtask difficulty).<sup>10</sup> In other words, each item in the grid remains in the same row in which it appeared in the previous instrument.

**Exhibit 12** is a sample of the letter sound identification subtask, the version designed for use by assessors; **Exhibit 13** is a sample student stimulus sheet for this same subtask.

---

<sup>10</sup> While reordering within rows will limit significant changes in subtask difficulty, it is still recommended to test for order effects whenever possible.

## Exhibit 12. Sample: Assessor protocol, letter sound identification (Icibemba language, Zambia)

Sub-test 2. LETTER SOUND IDENTIFICATION	Page 1	60 seconds																																																																																																																								
<p><b>Ili ipepala nalikwata ifilembo ifili mu alufabeti wa Cibemba. Nomba njebako ifiunda fya ifi filembo, ulande fyonse ifyo wiishibe. Ibukisha ukuti temashina yalefwaikwa iyoo, leelo fiunda.</b> Here is a page full of letters of the Cibemba alphabet. Please tell me the SOUNDS of as many letters of the alphabet as you can. Not their names, but their sounds.</p> <p>[point to the letter T] <b>Icilangililo, iciunda ca cilembo ici t, ni /t/</b> For example, the sound of this letter is /t/.</p> <p>[point to the letter M] <b>Natweshe ukucita ifi: Njebako iciunda ca cilembo ici:</b> Let's practice: Tell me the sound of this letter.</p> <p>✓ <b>Eya cawama, iciunda ca cilembo ici ni /m/.</b> Good, the sound of this letter is /m/.</p> <p>✗ <b>Iciunda ca cilembo ici ni /m/.</b> The sound of this letter is /m/.</p> <p>[point to the letter S] <b>Nomba natweshwa icilembo cimbi: Njebako iciunda ca cilembo ici.</b> Now let us try another one. Tell me the sound of this letter.</p> <p>✓ <b>Eya cisuma, iciunda ca cilembo ici ni /s/.</b> Good, the sound of this letter is /s/.</p> <p>✗ <b>Iciunda ca icilembo ici ni /s/.</b> The sound of this letter is /s/.</p> <p>[point to first letter] <b>Nganati "tampa", utampe mpaka upwishe ipepala lyonse. uleesonta pali cila cilembo na ukunjeba iciunda ca cilembo mu kwikatisha ishiwi. Ubelenge mukwangufyanya kabili busaka-busaka. Ngawasanga icilembo ushishibe, wikokolapo konkanyapo ukwabula ukupoosa inshita kabiye pa cilembo cakonkapo. Biika umunwe pa cilembo ca kubalilapo. Nauipekanya? Tampako.</b> When I say "Begin," start here and go across the page. Point to each letter and tell me the sound of that letter in a loud voice. Read as quickly and carefully as you can. If you come to a letter you do not know, go on to the next letter. Put your finger on the first letter. Ready? Begin.</p>	<p>Start the timer when the child reads the first letter.</p> <p>☞ If a child hesitates or stops on a letter for <u>3 SECONDS</u>, point to the next letter and say "Go on"</p> <p>👏 When the timer reaches 0, say "stop."</p> <p>👏 If the child does not provide a single correct response on the first line (10 items), say "Thank you!", discontinue this subtask, check the box at the bottom, and go on to the next subtask.</p>																																																																																																																									
<p>✗ (/) Mark any incorrect letters with a slash            (Ø) Circle self-corrections if you already marked the letter incorrect            ( ) Mark the final letter read with a bracket</p> <p><i>Examples:</i>    t    m    s</p> <table border="1" data-bbox="284 1131 1086 1512"> <thead> <tr> <th>1</th> <th>2</th> <th>3</th> <th>4</th> <th>5</th> <th>6</th> <th>7</th> <th>8</th> <th>9</th> <th>10</th> <th></th> </tr> </thead> <tbody> <tr> <td>e</td> <td>F</td> <td>u</td> <td>t</td> <td>W</td> <td>a</td> <td>p</td> <td>b</td> <td>L</td> <td>a</td> <td>(10)</td> </tr> <tr> <td>U</td> <td>a</td> <td>e</td> <td>s</td> <td>o</td> <td>i</td> <td>B</td> <td>k</td> <td>E</td> <td>A</td> <td>(20)</td> </tr> <tr> <td>N</td> <td>F</td> <td>P</td> <td>Y</td> <td>c</td> <td>a</td> <td>M</td> <td>I</td> <td>u</td> <td>L</td> <td>(30)</td> </tr> <tr> <td>i</td> <td>A</td> <td>K</td> <td>η</td> <td>a</td> <td>L</td> <td>i</td> <td>a</td> <td>s</td> <td>M</td> <td>(40)</td> </tr> <tr> <td>u</td> <td>t</td> <td>U</td> <td>K</td> <td>m</td> <td>o</td> <td>u</td> <td>n</td> <td>i</td> <td>A</td> <td>(50)</td> </tr> <tr> <td>b</td> <td>a</td> <td>n</td> <td>a</td> <td>E</td> <td>a</td> <td>O</td> <td>u</td> <td>s</td> <td>E</td> <td>(60)</td> </tr> <tr> <td>A</td> <td>n</td> <td>a</td> <td>S</td> <td>M</td> <td>L</td> <td>m</td> <td>η</td> <td>b</td> <td>T</td> <td>(70)</td> </tr> <tr> <td>u</td> <td>t</td> <td>i</td> <td>w</td> <td>I</td> <td>u</td> <td>B</td> <td>c</td> <td>N</td> <td>I</td> <td>(80)</td> </tr> <tr> <td>a</td> <td>I</td> <td>w</td> <td>a</td> <td>i</td> <td>N</td> <td>k</td> <td>m</td> <td>a</td> <td>L</td> <td>(90)</td> </tr> <tr> <td>y</td> <td>P</td> <td>M</td> <td>A</td> <td>U</td> <td>O</td> <td>A</td> <td>n</td> <td>a</td> <td>A</td> <td>(100)</td> </tr> </tbody> </table>	1	2	3	4	5	6	7	8	9	10		e	F	u	t	W	a	p	b	L	a	(10)	U	a	e	s	o	i	B	k	E	A	(20)	N	F	P	Y	c	a	M	I	u	L	(30)	i	A	K	η	a	L	i	a	s	M	(40)	u	t	U	K	m	o	u	n	i	A	(50)	b	a	n	a	E	a	O	u	s	E	(60)	A	n	a	S	M	L	m	η	b	T	(70)	u	t	i	w	I	u	B	c	N	I	(80)	a	I	w	a	i	N	k	m	a	L	(90)	y	P	M	A	U	O	A	n	a	A	(100)	
1	2	3	4	5	6	7	8	9	10																																																																																																																	
e	F	u	t	W	a	p	b	L	a	(10)																																																																																																																
U	a	e	s	o	i	B	k	E	A	(20)																																																																																																																
N	F	P	Y	c	a	M	I	u	L	(30)																																																																																																																
i	A	K	η	a	L	i	a	s	M	(40)																																																																																																																
u	t	U	K	m	o	u	n	i	A	(50)																																																																																																																
b	a	n	a	E	a	O	u	s	E	(60)																																																																																																																
A	n	a	S	M	L	m	η	b	T	(70)																																																																																																																
u	t	i	w	I	u	B	c	N	I	(80)																																																																																																																
a	I	w	a	i	N	k	m	a	L	(90)																																																																																																																
y	P	M	A	U	O	A	n	a	A	(100)																																																																																																																
<p>✗ Time remaining on stopwatch at completion (number of SECONDS)</p>																																																																																																																										
<p>✗ Exercise discontinued because the child had no correct answers in the first line</p>																																																																																																																										

**Eya cawama waesha! Katuleya ku cipande cakonkapo.** Good effort! Let's go on to the next section.

**Exhibit 13. Sample: Student stimulus sheet, letter sound identification (Icibemba language, Zambia)**

	t	m	s						
e	F	u	t	W	a	p	b	L	a
U	a	e	s	o	i	B	k	E	A
N	F	P	Y	c	a	M	l	u	L
i	A	K	ŋ	a	L	i	a	s	M
u	t	U	K	m	o	u	n	i	A
b	a	n	a	E	a	O	u	s	E
A	n	a	S	M	L	m	ŋ	b	T
u	t	i	w	l	u	B	c	N	l
a	l	w	a	i	N	k	m	a	L
y	P	M	A	U	O	A	n	a	A

**4.2.3 Nonword Reading**

Nonword reading is a measure of decoding ability (i.e., the sublexical route of word processing, as presented in Section 3.3.1) as distinct from whole word recognition or memorization, i.e., the lexical route. Many children in the early grades learn to memorize or recognize by sight a broad range of words. Exhaustion of this sight-word vocabulary at around age 10 has been associated with the “fourth-grade slump” in the United States (Hirsch, 2003). To be successful readers, children must combine both decoding and whole-word recognition skills; tests that do not include a decoding exercise can overestimate children’s ability to read unfamiliar words, as the words being tested may be part of the sight-recognition vocabulary.

**Data.** A child’s score is calculated as the number of correct nonwords per minute. The same categories of variables as collected for the other timed exercises are electronically collected for nonword reading: total correct words read, total incorrect words, and time remaining.

**Item construction.** This portion of the assessment includes a list of 50 one- and two-syllable nonwords, five per row, with the patterns of letters within the words adjusted as appropriate by language. Nonwords follow the rules of the language, using letters in legitimate positions (e.g., in English, not “wuj” because “j” is not used as a final letter in English). Also, they are restricted to consonant-vowel combinations that are typical of the language and are not homophones of real words (e.g., in English, not “kat,” homophone of “cat”). The grid uses a clear, well-spaced font. The items within rows of the grid can be reordered (re-randomized) for preparing equivalent test forms, although testing for ordering effects is recommended.

Exhibit 14 is a sample nonword reading subtask.

### Exhibit 14. Sample: Nonword reading (Icibemba language, Zambia)

Sub-test 3. NON-WORD READING	Page 2	60 seconds																																																																													
<p><b>🔊</b> Apa pali amashiwi aya kupangafye ayashilepilibula nangu cimo mu Cibemba. Ndefwaya ukuti ubelenge aya mashiwi yonse ayo wingabelenga. Wilalumbula ifilembo cimo-cimo iyoo kanofye ukubelenga ishiwi lyonse. Here are some made-up words in Icibemba. I would like you to read as many as you can. Do not spell the words, but read them.</p> <p>[point to the word “opa”] <b>Icilangililo: Ili shiwi lyapangwa ilyakuti: “opa”.</b> For example, this made-up word is: “opa”.</p> <p>[point to the word “toti”] <b>Natweshe nomba: belenga ili shiwi.</b> Let’s practice: Please read this word.</p> <p>✔🔊 <b>Eya cawama, ilishiwi ni “toti”.</b> Good, This made-up word is “toti.”</p> <p>✖🔊 <b>Ilishiwi lyakupangafye “toti” talipilibula nangu cimo.</b> This made-up word is “toti.”</p> <p>[point to the word “maba”] <b>Nomba esha nalimbi: Belenga nalimbi ishiwi ili.</b> Now let us try another one. Please read this word.</p> <p>✔🔊 <b>Ciisuma, ilishiwi lyaku pangafye ni “maba”.</b> Good, This made-up word is “maba.”</p> <p>✖🔊 <b>Ili shiwi lyaku pangafye ni “maba”.</b> This made-up word is “maba.”</p> <p>[point to first word] <b>Ilyo ndetila “Tampa” utampile apa no kubelenga yonse ayali pepepala lyonse. Uleesonta pali cila ishiwi na ukubelenga ukwikatisha ishiwi. Belenga mukwangufyanya kabili mu mutekatima. Ngawasanga ishiwi ushishibe wikokolapo uye palikonkelepo. Sonta peeshiwi lyaku balilapo. waipekanya? Tampako.</b> When I say “Begin,” start here [point to first word] and read across the page [point]. Point to each word and read it in a loud voice. Read as quickly and carefully as you can. If you come to a word you do not know, go on to the next word. Put your finger on the first word. Ready? Begin.</p>		<p>🕒 Start the timer when the child reads the first word.</p> <p>⏸️ If a child hesitates or stops on a letter for <u>3 SECONDS</u>, point to the next word and say “Go on”</p> <p>🕒 When the timer reaches 0, say “stop.”</p> <p>🕒 If the child does not provide a single correct response on the first line (5 items), say “Thank you!”, discontinue this subtask, check the box at the bottom, and go on to the next subtask.</p>																																																																													
<p>✖ (/) Mark any incorrect words with a slash            (Ø) Circle self-corrections if you already marked the word incorrect            ( ) Mark the final word read with a bracket</p> <p><i>Examples:</i> opa    toti    maba</p> <table border="1" style="width: 100%; text-align: center;"> <thead> <tr> <th></th> <th>1</th> <th>2</th> <th>3</th> <th>4</th> <th>5</th> <th></th> </tr> </thead> <tbody> <tr> <td>lebi</td> <td>ndite</td> <td>luti</td> <td>oya</td> <td>lusi</td> <td></td> <td>(10)</td> </tr> <tr> <td>mibu</td> <td>kibe</td> <td>shuti</td> <td>tobe</td> <td>njolo</td> <td></td> <td>(20)</td> </tr> <tr> <td>angi</td> <td>shipe</td> <td>nomi</td> <td>sani</td> <td>opu</td> <td></td> <td>(30)</td> </tr> <tr> <td>nepa</td> <td>wipi</td> <td>tupu</td> <td>naye</td> <td>koi</td> <td></td> <td>(40)</td> </tr> <tr> <td>tate</td> <td>shuma</td> <td>telu</td> <td>shingu</td> <td>yoba</td> <td></td> <td>(50)</td> </tr> <tr> <td>seni</td> <td>nupa</td> <td>etu</td> <td>kika</td> <td>onu</td> <td></td> <td>(60)</td> </tr> <tr> <td>sale</td> <td>pafu</td> <td>tawe</td> <td>ebi</td> <td>ewa</td> <td></td> <td>(70)</td> </tr> <tr> <td>ipa</td> <td>ombi</td> <td>kendi</td> <td>ngopa</td> <td>ndika</td> <td></td> <td>(80)</td> </tr> <tr> <td>afu</td> <td>yema</td> <td>mawe</td> <td>tebi</td> <td>folo</td> <td></td> <td>(90)</td> </tr> <tr> <td>fimu</td> <td>yapo</td> <td>tibu</td> <td>bife</td> <td>lefu</td> <td></td> <td>(100)</td> </tr> </tbody> </table>			1	2	3	4	5		lebi	ndite	luti	oya	lusi		(10)	mibu	kibe	shuti	tobe	njolo		(20)	angi	shipe	nomi	sani	opu		(30)	nepa	wipi	tupu	naye	koi		(40)	tate	shuma	telu	shingu	yoba		(50)	seni	nupa	etu	kika	onu		(60)	sale	pafu	tawe	ebi	ewa		(70)	ipa	ombi	kendi	ngopa	ndika		(80)	afu	yema	mawe	tebi	folo		(90)	fimu	yapo	tibu	bife	lefu		(100)	
	1	2	3	4	5																																																																										
lebi	ndite	luti	oya	lusi		(10)																																																																									
mibu	kibe	shuti	tobe	njolo		(20)																																																																									
angi	shipe	nomi	sani	opu		(30)																																																																									
nepa	wipi	tupu	naye	koi		(40)																																																																									
tate	shuma	telu	shingu	yoba		(50)																																																																									
seni	nupa	etu	kika	onu		(60)																																																																									
sale	pafu	tawe	ebi	ewa		(70)																																																																									
ipa	ombi	kendi	ngopa	ndika		(80)																																																																									
afu	yema	mawe	tebi	folo		(90)																																																																									
fimu	yapo	tibu	bife	lefu		(100)																																																																									
<p>🕒 Time remaining on stopwatch at completion (number of SECONDS)</p>																																																																															
<p>🕒 Exercise discontinued because the child had no correct answers in the first line</p>																																																																															
<p><b>Eya cawama waesha! Katuleya kucipande cakonkapo.</b> Good effort! Let’s go on to the next section.</p>																																																																															

#### 4.2.4 Oral Reading Passage with Comprehension

Oral reading fluency is a measure of overall reading competence: the ability to translate letters into sounds, unify sounds into words, process connections, relate text to meaning, and make inferences to fill in missing information (Hasbrouck & Tindal, 2006). As skilled readers translate text into spoken language, they combine these tasks in a seemingly effortless manner; because oral reading fluency captures this complex process, it can be used to characterize overall reading ability. Tests of oral reading fluency, as measured by timed assessments of correct words per minute, have been shown to have a strong correlation (0.91) with the Reading Comprehension subtest of the Stanford Achievement Test (Fuchs et al., 2001; Piper & Zuilkowski, 2015). Poor performance on a reading comprehension tool would suggest that the student may have trouble with decoding, or with reading fluently enough to comprehend, or with vocabulary.

**Data.** Students are scored on the number of correct words per minute and the number of comprehension questions answered acceptably. There will be two student scores: the number of words read correctly in the time allotted, and the proportion of questions correctly answered. The same three categories of variables collected for the other timed subtasks are electronically collected: total correct words read, total incorrect words, and time remaining. In addition, results for each of the comprehension questions are electronically recorded and entered into the database, with a final score variable calculated as a share of total questions asked. Data collection software prompts the assessor to ask only questions related to the text the child has read (see structure of questions and paragraph under “item construction” below).

**Item construction.** To create the oral reading fluency with comprehension subtask, the instrument developers review narratives from children’s reading materials. A narrative story has a beginning section where the characters are introduced, a middle section containing some dilemma, and an ending section with an action resolving the dilemma. It is not be a list of loosely connected sentences. The length of the story is about 60 words.

Character names frequently used in the school textbook are to be avoided, as students may give automated responses based on the stories with which they are familiar. However, character names must be typical of the language and context. Likewise, the story has only one to two characters, to avoid the task becoming about memory recall; and the names and places reflect the local culture.

The story text contains some complex vocabulary (inflected forms, derivations, etc.) and sentence structures. A large, clear, familiar font and good spacing between lines are used to facilitate student reading. No pictures are included.

The associated list of comprehension questions includes ones that can be answered directly from the text as well as at least one inferential question requiring students to combine knowledge and experience from outside the text to respond correctly. These inferential questions will have more than one right answer, but the answers must be logical based on the text and the context. Literal questions that are linked directly to the oral reading passage are the easiest type of comprehension measure. Including inferential questions in the subtask can provide insight into whether pupils are able to

connect the passage content with their own knowledge. The protocol for the subtask will specify the types of answers that may be marked as “correct.”

When equivalent forms of this subtask are to be created for use across multiple implementations of the same instrument in the same language (e.g., baseline, midterm, and endline), it is recommended to make simple changes in the story in order to limit the impact of test leakage, while retaining similar test difficulty. For example, names of story subjects, actions, and adjectives can be replaced with similar grade-level alternatives.

**Exhibit 15** is a sample of the oral reading fluency subtask for the Luvale language, including the reading comprehension component.

### Exhibit 15. Sample: Oral reading passage with reading comprehension (Luvale language, Zambia)

Sub-test 5a. ORAL READING PASSAGE	⌚60 seconds	Sub-test 5b: READING COMPREHENSION
<p>Show the child the sheet in the student stimulus booklet as you read the instructions.</p> <p>🔊 <b>Vanakuhane mujimbu wawih. Utangile helu, muwashiwashi, oloze mujila yakwoloka. Nge unakumisa kutanga, nangukuhulisa vihula hali ovyo unatange. Nge ngwami “putuka”, tanga mujimbu mujila yambwende nauhashilamo. Nge nauwana lizu lize kawatachikijileko, yako halizu likwavo. Haka munwe wove halizu lyatete. Unalizange? Putuka.</b> Here is a short story. I want you to read it aloud, quickly but carefully. When you finish, I will ask you some questions about what you have read. When I say “Begin,” read the story as best as you can. If you come to a word you do not know, go on to the next word. Put your finger on the first word. Ready? Begin.</p> <p>⌚ (/) Mark any incorrect letters with a slash            (Ø) Circle self-corrections if you already marked the letter incorrect            ( ) Mark the final letter read with a bracket</p>	<p>➡ If a child hesitates or stops on a letter for ≥ SECONDS, say “Go on”</p> <p>⌚ If the child does not provide a single correct word on the first line of text. Do not ask any comprehension questions.</p> <p>If a child says “I don’t know,” mark as incorrect.</p>	<p>After the child is finished reading, REMOVE the passage from in front of the child.</p> <p>Ask the child only the questions related to the text read. A child must read all the text that corresponds with a given question. If the child does not provide a response to a question after 10 seconds, mark “no response” and continue to the next question. Do not repeat the question.</p> <p>🔊 <b>Nangukuhulisa jino vihula vyavindende hali mujimbu unatange. Eseka kukumbulula vihula vize nauhasa. Unahase kukumbulula vihula mulilimi unasake.</b> Now I am going to ask you a few questions about the story you just read. Try to answer the questions as well as you can. You can provide your answers in whichever language you prefer.</p> <p>⌚ (✓) 1 = Correct            (✓) 0 = Incorrect            (✓) . = No response.</p>
		<b>Questions [Answers]</b>
Kwapwile lunga naphwevo vatwaminenga ukhawavo <u>mumusenge</u> .	6	<b>Ou lunga naphwevo vatwaminenga kuli?</b> (Mumusenge) 1 0 .
Vawanyinenga kukaluhwa mukuyoya chavo mwomwo kwakulimina kwapwile kwauchi. Kuyoya chavo chapwile chakukwata tuswa nakulya uchi. Echi chapwile nakuneha ushona kuli phwevo <u>chikuma</u> .	21	<b>Vayoyelenga hakulya ika?</b> (Tuswa nakulya uchi) 1 0 .
Lunga ashinganyekele akuya nakutunga kwakamwih navausoko <u>wenyi</u> .	28	<b>Iya ewwilenga ushona?</b> (Phwevo) 1 0 .
Chiyoyelo chavo chalumukile kaha vathu vevwile kuwaha <u>chikuma</u> .	35	<b>Chuma muka ashinganyekele lunga kulinga?</b> (Kuya nakutunga kuva usoko wenyi) 1 0 .
	43	<b>Chuma muka chanehesele vathu kuwahilla?</b> (Mwomwo chiyoyelo chavo chalumukile) 1 0 .
⌚ Time remaining on stopwatch at completion (number of SECONDS)		
⌚ Exercise discontinued: the child had no correct answers in the first line		
<b>Unazate kanawa, tutale jino vyuma vikwavo. Good effort! Let’s go on to the next section.</b>		

#### 4.2.5 Orientation to Print

Assessing a child’s knowledge of orientation to print can indicate that a child has been exposure to printed material in some way. Knowing concepts such as where to start reading and in which direction to read give provide evidence that a child has been given instruction about print material and its purpose. While research shows print awareness has little predictive ability regarding a child’s success with more advanced reading skills, it does have the ability to measure an overall literacy environment (Gove and Wetterberg, 2011). The subtask provides pupils with a short paragraph and asks basic questions—such as where to begin reading, or in which direction to read the text—in order to gauge pupils’ level of access (or not) to printed materials.

**Data.** The number correct out of the total number of questions asked is recorded. Additionally, it is often suggested to conduct item level analysis for this subtask.

**Item construction.** Provide the pupil with a short passage. It is important to instruct the child that no reading of the text is required. The pupil will simply use the text to demonstrate and answer the questions which are read aloud by the assessor. The questions asked by the assessors generally include the following (or some similar variation thereof):

1. On this page, where would you begin reading?
2. In which direction would you read the text on the page?
3. When you get to the end of the line, where would you read next?

The pupil answers the questions by pointing and demonstrating with his or her finger. The assessor marks each response as correct or incorrect. **Exhibit 16** is a sample of this subtask in Kikaonde.

### Exhibit 16. Sample: Orientation to print (Kikaonde language, Zambia)

Sub-test 4. ORIENTATION TO PRINT		Page 3	X
<p>☛ Show the child a story passage in the pupil stimuli packet. Read the instructions in the gray boxes below, recording the child's response before moving to the next instruction.</p>		Materials: a passage from the pupil stimuli packet	
<p><b>Keechi nkeba naamba utaange uno peepela aluno ine. Pa yino peeje. Waakoonsha kuteendekela peepi kutaanga? Toongolaapo na munwe woobe.</b></p> <p>I don't want you to read this now. On this page, where would you begin to read? Show me with your finger.</p>			
1. (Child puts finger on the top row, left-most word)	<input type="radio"/>	Correct	<input type="radio"/>
	<input type="radio"/>	Incorrect	<input type="radio"/>
	<input type="radio"/>	No Response	
<p><b>Pano mweesha ko usa kutazha kutaanga byaaloondelaapo.</b></p> <p>Now show me in which direction you would read next.</p>			
2. (Child moves finger from left to right)	<input type="radio"/>	Correct	<input type="radio"/>
	<input type="radio"/>	Incorrect	<input type="radio"/>
	<input type="radio"/>	No Response	
<p><b>Inge waafika kwaapela kipelu. Usa kutaanga byeepi jikwaabo?</b></p> <p>When you get to the end of the line, where would you read next?</p>			
3. (Child moves finger to left-most word of second line)	<input type="radio"/>	Correct	<input type="radio"/>
	<input type="radio"/>	Incorrect	<input type="radio"/>
	<input type="radio"/>	No Response	
Total Correct			/3

#### 4.2.6 English Vocabulary

Oral vocabulary tests are used for assessing a child in a language of instruction that differs from their first language. Children are asked a series 5 to 10 oral questions which have a child point to or demonstrate the answer. Body parts or basic classroom materials (i.e., pencil, paper, eraser) are often words that students are asked to identify. This subtask may also incorporate spatial vocabulary questions

such as under the paper, beside the paper, etc. These types of simple commands, given in a language of instruction, can indicate whether children possess basic vocabulary skills.

**Data.** The number correct out of the total number of words or phrases is recorded.

**Item construction.** Select 5 to 10 grade-appropriate vocabulary words that the student will be instructed to identify. The instructions will be read aloud by the assessor in the local language and only the actual vocabulary word(s) will be given in the language of instruction. Pictures of the words are typically avoided, and instead, students are asked to identify actual objects in front of them or body parts. After assessors provides the instructions asking the student to “point or show,” they read aloud the list of vocabulary words in the language of instruction one at a time while the student demonstrates their understanding of the word.

Spatial commands can also be incorporated via short phrases that instruct a student to place his or her pencil on, next to, or under a piece of paper. Again, the general instructions are given in the first language while the phrase (for example, “on the paper” or “under the paper”) is read in the language of instruction.

**Exhibit 17** is a sample of the English vocabulary subtask from the Zambia 2014 EGRA.

## Exhibit 17. Sample: English vocabulary knowledge (instructions in Lunda language)

Sub-test 6. ENGLISH VOCABULARY		Materials: a sheet of paper, pencil, rubber	⌚ X
<p>☞ (/) Mark any incorrect words with a slash (∅) Circle self-corrections if you already marked the word incorrect</p>			
<b>A. Body Parts:</b>			
<p>☞ <b>Nukutena mazhina ayiidi yamuzhimba muchizungu. Ntalishi chiidi chamuzhimba chinukutena. Twaya tudizi: “nose” Say, I’ll say words in English that represent parts of the body. Show me what part of your body the word means. Let’s practice.</b> “nose” (Point to your nose so that you model for the student) “head” Wait for the child to gesture to his/her head. Thereafter say, chachiwahi, komana wunachitiyi! Twaya tutachiki. Good you understand the directions! Let’s start.</p>			
<p>_____</p> <p>foot          arm          chin          knee          ear          back          elbow          shoulder</p>			
Part A Total Correct			/8
<b>B. Words from the Environment:</b>			
<p>☞ <b>Ichi nukutena mazu amakwawu dichi eyi wukuntalisha Yuma yinakutena owu mazu.</b> Now I will say other words and you will show me examples of those words.</p>			
<p>pencil      shoes      desk      rubber      paper      floor</p>			
Part B Total Correct			/6
<b>C. Spatial Words</b>			
<p>☞ <b>Tambulaku pensulu iyi. (Hand the pencil to the child.) Wukuyisha oyu pensulu ohu hinukukulezha kuyisha. Shaku pensulu ha .... Say, Take this pencil. You will place the pencil where I tell you to put it. Put the pencil...</b></p>		Place a pencil and sheet of paper side by side in front of the student.	
<p>on the paper      next to the paper      behind you      under the paper      in front of you      to the right of the paper</p>			
Part C Total Correct			/6
Overall Total Correct = (Part A + Part B + Part C)			/20

### 4.3 Translation and Other Language Considerations

#### 4.3.1 Translation vs. Adaptation

The consensus among education experts is that when evaluators are developing or modifying EGRA instruments, it is not viable to simply translate either the words or

the connected-text passage from a version in a different language. Quite simply, translation may result in use of inappropriate words in the mother tongue that are too difficult for the grade level. For example, translating a syllable-segmenting task from English to Spanish when the word being segmented is “yesterday” would result in comparing a three-syllable word with a two-syllable word (“ayer” in Spanish), which would reduce the reliability of the assessment instrument and the validity of the cross-linguistic comparisons of results. As discussed earlier in this section, careful work in an adaptation workshop results in original passages that are approximately equal in difficulty to the texts students are expected to read at grade level in each context.

The instructions must be translated as closely as possible to the original EGRA instructions, capturing the meaning more than a verbatim version.

Noted early in EGRA’s development by Penelope Collins (née Chiappe) in a 2006 personal communication relating her experience within the South Africa Department of Education,

Because of linguistic differences (orthographic and morphological), it is critical that the passages used are independently written. Equivalence between passages cannot be established by translating the English passage into the different languages.

This was clearly illustrated by the initial pilot of the isiZulu passage. The isiZulu passage was a translation of the English passage. Although one would expect children’s oral reading rate to be similar for the context-free word/nonword lists and the passage, isiZulu learners who could read 20–30 correct words per minute in the list could not read the passage at all. Closer inspection of the isiZulu passage revealed that the isiZulu words were much longer than those in the isiZulu list and the words used in the English passage. Thus, the isiZulu passage was clearly too difficult for students reading at a first-grade level.

*English:* “John had a little dog. The little dog was fat. One day John and the dog went out to play. The little dog got lost. But after a while the dog came back. John took the dog home. When they got home John gave the dog a big bone. The little dog was happy so he slept. John also went to sleep.”

*isiZulu:* “USipho wayenenja encane. Inja yakhe yayikhuluphele. Ngolunye usuku uSipho wayehamba nenja yakhe ukuyodlala. Inja yalahleka. Emva kwesikhathi inja yabuya. USipho waphindela ekhaya nenja yakhe. Emva kokufika ekhaya, uSipho wapha inja ekhaya ukudla okuningi. Inja yajabula kakhulu yaze yagcina ilele. NoSipho ngokunjalo wagcina elele.”

#### 4.3.2 Cross-Language Comparisons: Preparations and Considerations

The issue of comparability across languages and countries is challenging from an assessment perspective. EGRAs administered in different contexts or in different

languages may use comparable test forms meaning the tests are intended to be judged in relationship to each other and thus are designed with the same constructs, subtasks, etc. That is, the forms themselves have the same measurement purpose; however, there is no assumption of equivalence (i.e., identical item difficulty).

Research indicates the difference between languages may be primarily a matter of the *rate* at which the children achieve the first few steps toward reading acquisition (Seymour et al., 2003). Regardless of language, all children who learn to read advance from being nonreaders (unable to read words) to partial readers (can read some items but not others) to readers (can read all or a majority of items). In languages with transparent or “shallow” orthographies (often called phonetically spelled languages), the progression through these levels is very rapid (just a few months of learning); in languages with more complex or “deeper” orthographies, this process can take several years. In English, for example, completing the foundation steps requires two or more years, with a rate gain of only a few new items per month of learning. In comparison, regular and transparent languages such as Italian, Finnish, and Greek require only about a year of instruction for students to reach a comparable level (Seymour et al., 2003).

As languages have different levels of orthographic transparency, it is not easy to say that Country A (in which all children are reading with automaticity by grade 2) is outperforming Country B (where children reach this level only by grade 3), if Country A’s language has a far more transparent orthography than Country B’s language.

Nonetheless, finding out at which grade children are typically “breaking through” to reading in various countries, for example, and comparing these grades, could be a useful analytical and policy exercise. The need for this type of “actionable data” was one rationale behind the creation of the Early Grade Reading Barometer (<http://www.earlygradereadingbarometer.org/users/login>), an interactive tool developed with USAID funding. It uses actual EGRA data sets from dozens of countries to generate graphical displays of students’ reading performance, by country, and is publicly available (free login required).

In order to make reasonable cross-linguistic comparisons, educators and policy makers must complete two steps.

First, to ensure the technical adequacy<sup>11</sup> of an EGRA instrument across languages specifically, one must adapt, rather than translate, the instrument to account for differences in the cultural or linguistic elements of a language (as explained in Section 4.3.1 above). Even so, directly comparing all EGRA subtask results from one language’s assessment to another is not advised.

Second, in the case that comparison across languages is desired, those adapting and analyzing the EGRA results must, at a minimum, conduct a thoughtful examination of:

1. The technical adequacy of an assessment for its stated purpose;
2. The features of the languages, such as orthographic depth or orthographic complexity;

---

<sup>11</sup> A “technically adequate instrument” is one that has been demonstrated to produce reliable results, allows the generation of valid analyses, and therefore lends confidence.

3. Each subtask, to understand the overall and particular constructs they are attempting to capture.

For further guidelines and recommendations on how to adapt and compare EGRA results across languages, see **Annex A**.

#### **4.4 Using Same-Language Instruments Across Multiple Applications: Creation of Equivalent Test Forms**

As mentioned earlier in this section, adaptation can involve modifying an existing instrument that was previously developed for a given language. If there is no concern about test leakage (i.e., if teachers have limited access to EGRA instruments and it is unlikely that students will become familiar with a particular form of the assessment), the same instrument can simply be used across multiple time points. If however, leakage is a concern, it will be necessary to have multiple assessments (or test forms) that are used to measure changes in performance. In order to ensure that valid comparisons of results can be made across assessment forms/administrations, instruments must be modified in such a way as to create new forms that are as equal as possible in difficulty to the original form. Equivalent tests forms refers to tests that are intended to be of equal difficulty (and thus directly substitutable for one another).

It is true that in instances where subtask difficulty from EGRA instrument A and instrument B is determined post-test not to be equal, specific test equating procedures can be applied to account for the differences (see Section 8.5). Equated test forms, therefore, refers to forms that have been adjusted by a statistical process in order to make scores comparable. However, best practice for instrument and subtask modification recommends limiting the need for post-administration statistical equating. Techniques for preparing equivalent forms are described throughout the adaptation section of the toolkit (Section 4), and may include:

- Making simple changes in the names of story subjects, actions, and adjectives, replacing them with grade-level equivalents
- For subtasks that are presented to learners on stimuli sheets that are in a grid format, shuffling items within the grid rows.

For situations in which these techniques are used but still result in non-equivalent test forms, statistical equating methods may be required. Section 8.5 discusses specific methodologies and recommendations for equating scores after data are processed and analyzed.

#### **4.5 Best Practices**

As EGRA has expanded into dozens of countries and even more languages, many lessons have been learned that are worth bearing in mind in the planning and execution of both adaptation development and adaptation modification.

- **Instructions.** Debating the EGRA protocol, or the instructions the assessors are to follow, is unproductive. The instructions were carefully developed based on evidence from prior research and experience and are never modified. Instead, time spent on accurate translation of the instructions is critical for successful implementation.

- **Pretesting and pilot testing.** Both of these steps are important parts of the process (see first part of Section 4 as well as Section 7 of the toolkit) and must be planned and budgeted.
- **Minimum content.** At a minimum, an EGRA must test listening comprehension, letter sounds, nonword reading, and oral reading fluency with comprehension; other subtasks depend on contextual factors.
- **Use of the same or nearly identical subtask items across multiple forms of an instrument.** Best practice is to limit the need for post-administration statistical equating whenever possible. Strong instrument design procedures can produce highly comparable forms that mitigate the need for equating.

## 5 USING ELECTRONIC DATA COLLECTION

Starting in 2010, EGRA researchers began to transition from paper-based data collection to electronic data collection. Electronic data collection reduces the potential for errors or omissions in the data and makes results available more rapidly.

Comparisons of electronic versus paper-based data collection have shown advantages in terms of effectiveness and efficiency. The increasing availability of affordable mobile devices and Internet connectivity that allow researchers to analyze data in real time continue to drive support for e-data capture (Walther et al., 2011).

A key difference between electronic and paper-based data collection is the elimination of manual data entry of completed paper forms into an electronic

database. This reduces the time spent and potential errors associated with manual data entry from paper, as well as errors that result from assessors incorrectly or illegibly marking paper forms or skipping questions. Moreover, electronic data collection results can be uploaded from the field, and can be processed

*Electronic data collection improves and strengthens fieldwork.*

and analyzed sooner. This feature also provides an opportunity to detect and rectify issues while assessors are still in the field. Electronic data collection therefore improves and strengthens fieldwork.

It is important to keep in mind that electronic data collection does not change the basic implementation procedures of the assessment. The child still reads from a sheet of paper with the letters and words printed on it; the assessor still provides the same instructions. The instructions for electronic data collection do not change except in reference to how to mark responses (e.g., “mark” versus “touch the screen”).

The first known examples of wireless mobile data collection designed specifically for EGRA were iProSurveyor, developed by Prodigy Systems for use in Arabic in Yemen and then Morocco, in 2011;<sup>12</sup> and the software system Tangerine®, created by RTI International beginning in 2010 and piloted in 2012. These two software programs adapted the EGRA instrument, including timed tasks, to a discrete, portable, and intuitive touch-screen tablet interface that would not interfere with the basic one-on-one administration procedure of EGRA. The iProSurveyor EGRA effort in Yemen involved 38 schools in three governorates, with 735 student interviews in grades 2 and 3. Tangerine was first field-tested in January 2012 under the USAID Primary

<sup>12</sup> Under a subcontract to RTI International on the USAID EdData II project (see Collins & Messaoud-Galusi, 2012; Prodigy Systems, 2011).

Math and Reading (PRIMR) Initiative in Kenya, for which 176,000 data points were captured through a small sample of 200 pupils from 10 schools being assessed with an English EGRA, Kiswahili EGRA, and EGMA (Strigel, 2012). These field tests demonstrated ease of use and efficiencies gained, and electronic data collection was confirmed as a feasible approach to supersede paper data collection for oral reading (and math) assessments with timed components.

## 5.1 Cautions and Limitations to Electronic Data Collection

For electronic data collection, limitations to be aware of are:

- **Risk for error.** Electronic data collection is not foolproof. There is some degree of potential for input errors or loss of data.
- **Cost considerations.** Cost analyses carried out for USAID under EdData II have indicated that efficiencies of using electronic data collection over paper instruments are most commonly achieved when the hardware is used for multiple data collections. Cost savings may not occur if the required hardware is used only for a single data collection.
- **Need for paper backups.** Assessment teams still must carry some backup paper instruments in case the electronic hardware should fail while they are conducting the fieldwork. Therefore, paper instruments are introduced during assessor training along with the electronic software.
- **Limited exposure to technology.** Planners must take into account both the country/regional context and assessors' familiarity with technology when considering electronic data collection.
- **Security issues.** Loss, theft, and damage to devices create the potential for financial loss or personal harm, so ensuring the safety and security of the hardware and assessors necessitates careful planning.
- **Limited communications infrastructure.** Finding or creating remote, mobile hotspots for uploading field data can be difficult in some countries or regions.
- **Limited local capacity.** Adaptations of the instrument into local languages and scripts, and rendering the content into the chosen data collection software, present related challenges. Affiliations with experienced local partners are key in fully exploring and mitigating capacity limitations regarding e-data capture.

When using electronic data collection over paper data collection, researchers must also address the need to maintain the security of digital data; depending on the software used to collect the data, access to raw results may be accessible by multiple people. Even GPS points must be used only for verification purposes, and not to identify individual schools. As with paper-based research, every effort has to be taken to ensure that privacy is respected and that no individual schools, teachers, or students could be subjected to negative repercussions because of the results.

## 5.2 Data Collection Software

Many mobile survey tools exist that can be adapted for EGRA administration. The open-source program Tangerine is one widely used tool, applied in more than 60 implementations in 36 countries by 27 organizations as of mid-2015 (see

[www.tangerinecentral.org](http://www.tangerinecentral.org)). As of this writing, iProSurveyor (for the iPad), Tangerine, and SurveyToGo were the only platforms not including laptop or desktop data entry systems known to have been adapted to the EGRA. Implementers consider which software is most compatible with the context and the nature of the data being collected—in particular, the unique timed grid format of many EGRA subtasks and the need to calculate total number of items attempted (accuracy) and items correct per minute (fluency). Where the data are to be stored, who will manage it, and technical capacity may also be considerations in choosing particular software.

### **5.3 Considerations for Hardware Selection and Purchasing**

When procuring hardware to accommodate electronic EGRA data collection, implementers have to consider factors such as shipping, storage, and reuse of the materials. As of 2015, tablet computers (rather than mobile phones, smartphones, or laptops) are considered the most appropriate type of hardware because of screen size, ease of use, light weight, and especially, long battery life. At a minimum, additional accessories must include a stylus, protective case, and wireless router for effective data collection and ability to send results daily.

Implementers must weigh the pros and cons of purchasing hardware in the country where data collection will take place or purchasing outside of the country of implementation. External purchases will require planning sufficient lead time to account for shipping and clearing customs. Hand-carrying devices from one country to another is possible, in cases where only a small number of tablets and accessories are being used (or reused), but individuals carrying the hardware have to be aware of customs regulations and potential fees for importing devices, depending on local context. For example, some countries require proof of plans to export the devices after data collection before they will waive import duties.

Implementers must also plan for appropriate storage of all hardware and accessories before and after data collection, and during training. All devices and peripherals are required to be stored in a location that can be secured to deter theft. The storage area also should be protected from dust, humidity, and extreme temperatures. Note that battery life of devices can be affected after long periods of nonuse.

### **5.4 Supplies Needed for Electronic Data Collection and Training**

- Tablets, each with charger
- Software containing electronic version of assessment
- Tablet cases
- Styluses
- Bags for assessors to carry tablets to the field sites
- Hotspot routers and connectivity dongles plus a data plan
- Several extra tablets in case of damage or loss

# 6 EGRA ASSESSOR TRAINING

This section provides guidance on planning for and conducting an EGRA assessor training.

Note that this section is not intended to be an assessor or supervisor manual; rather, it is a resource for the training organizers. The *Guidance Notes for Planning and Implementing Early Grade Reading Assessments* contain additional details on assessor training and are recommended as a companion to this document (RTI International & International Rescue Committee, 2011).<sup>13</sup>

The assessors who will be piloting the instrument will need a training of about five working days.<sup>14</sup> The length will depend on factors such as the number of instruments to be administered (e.g., a mathematics assessment in addition to EGRA), the number of trainers available, the number of people to be trained, trainees' prior experience, and the budget and time available. For example, if some trainees will have limited proficiency in the language of the training (such that a translator may be required), it is wise to add two or three days to the schedule.

To ensure that all trainees understand the purpose of and endorse the work, a key element of the agenda will be reviewing the underlying EGRA principles and the reasoning behind the instrument components. Other main objectives are:

- To train a cohort of assessors to accurately and effectively administer the EGRA, in electronic and paper formats;
- To identify skilled individuals to serve as assessors for the data collection;
- To identify and train selected individuals to serve as supervisors during data collection.

## 6.1 Recruitment of Training Participants

It is vital to recruit and train 10% to 20% more assessors than the sampling plan indicates will be needed. Inevitably, some will not meet the selection criteria, and others may drop out after the training for personal or other reasons.

Data collection teams may be composed of education officials and/or independent assessors recruited for the particular data collection. Requirements and preferences are determined during the recruitment phase, in advance of the training, depending on the specific circumstances and purposes.

Government officials can be considered as candidates for the assessor or supervisor roles. In order to be selected for the fieldwork, however, they will need to meet the

---

<sup>13</sup> The Guidance Notes can be found on the EdData II website: <https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=357>

<sup>14</sup> See Section 7.1.3 on the pros and cons regarding the various possible timings of the assessor training in relation to the pilot and full data collection.

same performance standards as all other trainees. The facilitators must emphasize the selection standards at the beginning of the training. A potential benefit of involving qualified government officials is the greater likelihood of the government's positive reception to the data analysis once the results are announced. Important criteria for planners to consider when identifying people to attend the assessor training are the candidates'

- Ability to fluently read and speak the languages required for training and EGRA administration;
- Previous experience administering assessments or serving as a data collector;
- Experience working with primary-age children;
- Availability during the data collection phase and ability to work in target areas;
- Experience and proficiency using a computer or hand-held electronic device (tablet, smartphone).

The training team will select the final roster of assessors based on the following criteria. These prerequisites are communicated to trainees at the outset so they understand that final selection will be based on who is best suited for the job.

- **Ability to accurately and efficiently administer EGRA.** All those selected to serve as assessors must demonstrate a high degree of skill in administering EGRA. This includes knowledge of administration rules and procedures, ability to accurately record pupils' responses, and ability to use all required materials—such as a tablet—to administer the assessment. Assessors must be able to manage multiple tasks at once, including listening to the student, scoring the results, and operating a tablet.
- **Ability to establish a positive rapport with pupils.** It is important that assessors be able to interact in a nonthreatening manner with young children. Establishing a positive, warm rapport with students helps them to perform to the best of their abilities. While this aspect of test administration can be learned, not all assessors will master it.
- **Ability to work well as a team in a school environment.** Assessors do not work alone, but rather as part of team. As such, they need to demonstrate an ability to work well with others to accomplish all the tasks during a school visit. Moreover, they need to show they can work well in a school environment, which requires following certain protocols, respecting school personnel and property, and interacting appropriately with students.
- **Availability and adaptability.** As stated above, assessors must be available throughout the data collection, and demonstrate their ability to function in the designated field sites. For example, they may have to spend a week in a rural environment where transportation is challenging and accommodations are minimal.

From among the trainees, the facilitators also identify supervisors to support and coordinate the assessors during data collection. Supervisors (who may also be known as data collection coordinators, or other similar title) must meet if not exceed the criteria for assessors. In addition, they must:

- Exhibit leadership skills, have experience effectively leading a team, and garner the respect of colleagues.

- Be organized and detail-oriented.
- Know EGRA administration procedures well enough to supervise others and check for mistakes in data collection.
- Possess sufficient knowledge/skills of tablet devices in order to help others.
- Interact in an appropriate manner with school officials and children.

The facilitators must also communicate these qualifications in advance to trainees and any in-country data collection partners. Supervisors will not necessarily be people with high-level positions in the government, or those with another form of seniority. Officials who do not meet the criteria may be able to serve another supervisory role, such as drop-in site visits. Such situations sometimes arise when education officials would like to play some role in observing and supervising the data collection, whether or not they could attend the assessor training; benefits of accommodating them can be a greater understanding of the EGRA process and acceptance of the results.

## 6.2 Planning the Training Event

Key tasks that need to take place before the training event include:

- **Prepare EGRA instrument and training materials.** Finalize the content of the instruments that will be used during training—both electronic and paper, for all languages. Other training documents and handouts (e.g., agenda, paper copies of questionnaires and stimulus sheets, supervisor manual) also need to be prepared and copies made.
- **Procure equipment.** Materials and equipment that the planners anticipate and procure well in advance range from the tablets and cases, to flipchart paper, stopwatches, power strips, and pupil gifts. Create an inventory to keep track of all materials throughout the EGRA training and data collection.
- **Prepare equipment.** For those supporting the technology aspects of the training, once the tablets have been procured, they must be prepared for data collection. This means loading the software and electronic versions of the instruments onto the tablets and setting them up appropriately.
- **Prepare workshop agenda.** Create a draft agenda and circulate it among the team implementing the workshop. For an EGRA-only training, the main content areas in the agenda will include:
  - Overview of EGRA instrument (purpose and skills measured)
  - Administration of EGRA subtasks (protocols and processes; repeated practice)
  - Tablet use (functionality, saving and uploading of assessments)
  - Sampling and fieldwork protocols.

See **Annex B** for a sample agenda.

- **Finalize the facilitation team.** Assessor trainings are facilitated by at least two trainers who are knowledgeable about reading assessment (and EGRA in particular), and who have experience training data collectors. The trainers do not necessarily need to speak the language being tested in the EGRA instrument if

they are supported by a local-language expert who can verify correct pronunciation of letters and words, and assist with any translation that may be needed to facilitate the training. However, the trainers must be fluent in the language in which the workshop will primarily be conducted. If the training will be led in multiple languages, a skilled team of trainers is preferred and additional trainers can be considered.

### 6.3 Components of Assessor Training

As indicated via the sample agenda in Annex B, the assessor training will incorporate several consistent components. In a sequence similar to the following, the facilitators:

- Invite high-level officials whose purpose is to publicly state their commitment to the EGRA and their interest in the results.
- Introduce the assessment project, the importance of early grade reading, what the EGRA is, and the basics of instrument administration.
- Explain the importance to the research of monitoring the assessors' performance, and the criteria by which they will be evaluated and selected.
- Give an overview of the subtasks; demonstrate how they are administered.
- Present and explain any supplemental instruments to be administered alongside the EGRA.
- Give the participants opportunities to practice in pairs and groups, with oversight and support from the lead trainers. After several days of training, arrange for at least one practice with children in a school setting.
- Observe, assist, and retrain as needed. Ensure that the trainees become comfortable with both the survey content and the equipment and software.
- Formally evaluate assessor accuracy (refer to Section 6.7); use the results for remediation and ultimately for selecting the assessor corps for the main data collection.

### 6.4 Training Methods and Activities

Research on adult learning points to some best practices that should be employed in an assessor training. Whether the training involves a team of 20 assessors or 100, creating *interactive sessions* in which participants work with each other, the technology, and instrument will result in more effective learning.

Experience training EGRA assessors globally indicates that the more opportunities participants have to practice EGRA administration, the better they learn to effectively administer the instrument. In addition, *varying activities* from day to day will allow participants the opportunity for deeper engagement and better outcomes. For example, day-to-day activities for training on the tablet can include:

- Facilitator demonstrations
- Videos
- Whole-group practice
- Small-group practice

- Pairs practice
- Trainee demonstrations

Throughout the training, facilitators should vary the pairs and small groups. This may include pairing a more skilled or experienced assessor with someone less experienced.

Some ideas include a “round-robin” approach to practicing items that need the most review (e.g., participants sit in a circle and take turns quickly saying the sounds of the letters in the EGRA instrument); or simulations in which a person playing the role of an assessor makes mistakes or does not follow proper procedures, then participants are asked to discuss what happened and what the “assessor” should have done differently.

If more than one language will be involved, it is advised to keep these activities within the language groups.

The facilitators will need to direct the trainees to also spend time practicing tablet functionality: drop-down menus, unique input features, etc.

Showing workshop participants videos of the EGRA being administered can help them to understand the process and protocols before they have an opportunity to administer it themselves. These videos—which will require appropriate permissions and will need to be recorded in advance of the training—can be used to model best practices and frequently encountered scenarios. They can serve as a useful springboard for discussions and practice.

## 6.5 School Visits

Assessor training always involves, at a minimum, one school visit to allow assessors to practice administering the EGRA to children and using the technology in conditions similar to those they will encounter during actual data collection. The school visits also allow them to practice pupil sampling procedures and to complete all required documentation about the school visit.

To help ensure productive school visits, the training leadership team will:

- Schedule at least one school visit during training (two or more would be preferable):
  - Plan for one halfway through the training, and one toward the end.
- Identify how many schools are needed:
  - Base the number of schools on the number of trainees, size of nearby schools, number of visits.
  - Avoid overwhelming schools by bringing too many people to one school. Assign no more than 35–40 people to a large school but fewer for smaller schools.
- Identify schools in advance of the training:
  - Get required permission, alert principals, and plan for transportation; verify schools are not part of the full data collection sample (if this is not

possible, make sure to exclude the practice schools from the final sample).

- Prepare teams a day in advance so they know what to expect:
  - Departure logistics, who's going where, team supervisors, number of students per assessor, assessments to be conducted, etc.

## SUMMARY OF TRAINERS' DUTIES DURING SCHOOL PRACTICE VISITS

- Identify trainees to serve as supervisors
- Help teams with introductions as needed
- Observe assessors and provide assistance as needed
- With appropriate permission: Take photos or videos of the assessors, for further training and discussion during debrief
- Return classrooms/resources to the way they were when the teams arrived
- Thank the principal for time and participation



*In-school practice of EGRA administration during the 2016 G2 NAS Planning Workshop held in February 2016 in Livingstone, Zambia (Credit: RTI staff)*

A quiet and separate space at the school will be needed for participants to practice administering the assessments. As seen in the picture above, ideally, assessors should be able to sit across a desk from a child and administer the instrument. If

desks are not available, the child can sit in a chair that is placed at a slight diagonal from the assessor.

During the first school visit, it is helpful for participants to conduct the EGRA in pairs, so that they can observe and provide feedback to each other. Working in pairs is also helpful since participants are often nervous the first time they conduct an EGRA with a child.

During a second or third visit, participants may be more comfortable working on their own and will benefit from practicing administration with as many children as possible during the visit. They will also be able to practice pupil sampling procedures and other aspects of the data collection they may not yet have learned about before the first school visit.

Each assessor will administer the instrument(s) to between four and eight<sup>15</sup> children, each, at every school visit.

It is critically important after the visit to carry out a debriefing with the participants. It gives trainees an opportunity to share with the group what they felt went well, and what they found challenging. Often the school visit raises new issues and provides an opportunity to answer questions that may have come up during the training.

## **6.6 Assessor-Trainee Evaluation Process**

A transparent evaluation process and clear criteria for evaluation are helpful for both facilitators and trainees. The process used to evaluate assessors during training includes both formal and informal methods of evaluation. As part of the informal evaluation, facilitators observe trainees carefully during the workshop and school visits and also conduct one-on-one interviews with them, when possible.

Trainees will require feedback on both their strengths and challenges throughout the workshop. Having a qualified and adequate team of trainers will ensure that feedback is regular and specific. Likewise, having enough trainers will allow for feedback that addresses trainees' need for additional assistance, and for the careful selection of supervisors.

Careful observation of the assessors supports the collection of high-quality data—the ultimate goal. Therefore, whenever the assessors are practicing, facilitators are walking around monitoring and taking note of any issues that need to be addressed with the whole group.

Evaluation of assessors is multifaceted and takes into consideration several factors, among them the ability to:

- Correctly and efficiently administer instruments, including knowing and following all administration rules
- Accurately record demographic data and responses
- Identify responses as correct and incorrect

---

<sup>15</sup> The number of pupils each data collector is able to assess at a school depends heavily on the number of subtasks per instrument and the total number of instruments being administered.

- Correctly and efficiently use equipment, especially tablets
- Work well as a part of a team
- Adhere to school visit protocols
- Create a rapport with pupils and school personnel.

Throughout the training, participants themselves reflect on and share their experiences using the instrument. The training leaders are prepared to improve and clarify the EGRA protocol (i.e., the embedded instructions) based on the experience of the assessors both in the workshop venue and during school visits.

Formal evaluation of assessors has become standard practice in many donor-funded projects and is an expected outcome of an assessor training program. The next section goes into detail about measuring assessors' accuracy. Trainers evaluate the degree of agreement among multiple raters (i.e., assessors) administering the same test at the same time to the same student. This type of test or measurement of assessors' skills determines the trainees' ability to accurately administer the EGRA.

## 6.7 Measuring Assessors' Accuracy

As part of the assessor selection process, workshop leaders measure assessors' accuracy during the training by evaluating the degree to which the assessors agree in their scoring of the same observation.

### OVERVIEW OF FORMAL EVALUATION FOR MEASURING ASSESSORS' ACCURACY DURING TRAINING

1. **Assessing and selecting assessors.** Establish a benchmark. Assessors unable to achieve the benchmark are not selected for data collection. In an EGRA training, the benchmark is set at 90% agreement with the correct evaluation of the child for the final training assessment.
2. **Determining priorities for training.** These formal assessments indicate subtasks and items that are challenging for the assessors, which also constitute important areas of improvement for the training to focus on.
3. **Reporting on the preparedness of the assessors.** An assessor training involves three formal evaluations of assessors to assess and monitor progress of accuracy.

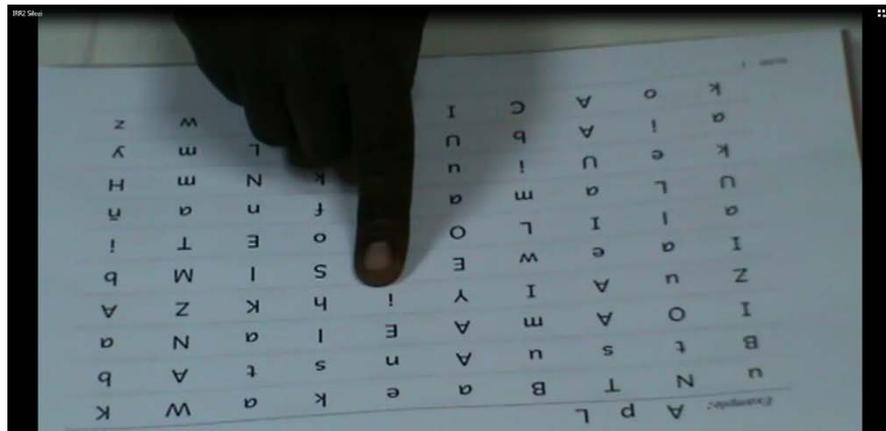
This type of evaluation is particularly helpful for improving the assessors' performance before they get to the field. It must also be used for selecting the best-performing assessors for the final assessor corps for the full data collection, as well as alternates and supervisors.

The training team creates a separate instrument in the tablets for the purpose of conducting the assessor accuracy measure.

There are two primary ways to generate data for calculating assessor accuracy:

1. If the training leaders were able to obtain appropriate permissions before the workshop and to make audio or video recordings of students participating in practice or pilot assessments (see **Exhibit 18**), then in a group setting, the recordings can be played while **all** assessors score the assessment as they would during a “real” EGRA administration. A skilled EGRA assessor also scores the assessment and those results are used as the Gold Standard.

### Exhibit 18. Frame from video used for assessment



2. Adult trainers or assessors can play the “student” and “assessor” roles in large-group settings (or on video) and assessors all score the activity. The benefit of this latter scenario is that the adults can deliberately and unambiguously make several errors on any given subtask (e.g., skipping or repeating words or lines, varying voice volume, pausing for extended lengths of time to elicit prompts, etc.). The script prepared beforehand, complete with the deliberate errors, becomes the Gold Standard.

The trainers will then upload all the trainees’ assessments into Excel or other analysis software and comparatively analyze the results. Refer to **Annex C** for more about data analysis and statistical guidance for measuring assessor accuracy.

After an assessor evaluation, the data need to be reduced to just the trainees’ attempts during the assessment along with the Gold Standard assessment.

If for some reason the training team did not create a Gold Standard before or during the trainees’ assessment, the lead trainer prepares one afterward and adds its results to the database. Additionally, the training team must review the Gold Standard responses to ensure that what is recorded for each Gold Standard response accurately reflects the consensus on the correct responses to the assessment. One important approach is to compare the Gold Standard with the mode (most frequent) response of the assessors at the item level.

As previously mentioned, measuring assessors’ accuracy is important as it helps a trainer identify assessors whose scoring results are *greater than one standard deviation* from the Gold Standard and who may require additional practice or support.

It can also be used to determine whether the entire group needs further review or retraining on some subtasks, or whether certain skills (such as early stops) need additional practice.

If the analysis from the formal evaluation reveals consistent poor performance on the part of a given assessor, and if performance does not improve following additional practice and support, that assessor cannot participate in the fieldwork. Again, refer to Annex C for more information about how to evaluate the assessor accuracy data.

# 7 FIELD DATA COLLECTION: PILOT TEST AND FULL STUDY

## 7.1 Conducting a Pilot EGRA

A pilot test is a small-scale preliminary study conducted prior to a full-scale survey. Pilot studies are used to conduct item-level assessments to evaluate each subtask as well as test the validity and reliability of the EGRA instrument and any accompanying questionnaires. Additionally, pilots can test logistics of implementing the study (cost, time, efficient procedures, and potential complications) and allow the personnel who will be implementing the full study to practice administration in an actual field setting.

In terms of evaluating the instruments that will be used during the data collection, the pilot test can ensure that the content included in the assessment is appropriate for the target population (e.g., culturally and age appropriate, clearly worded). It also is a chance to make sure there are no typographical errors, translation mistakes, or unclear instructions that need to be addressed.

### WHY CONDUCT A PILOT TEST OF THE EGRA?

A pilot test is used to

- Ensure reliability and validity of the instrument through psychometric analysis.
- Obtain data on multiple forms of the instruments, for equating purposes.<sup>16</sup>
- Review data collection procedures, such as the functionality of the tablets and e-instruments along with the procedures for uploading data from the field.
- Review the readiness of the materials.
- Review logistical procedures, including transportation and communication, among assessor teams, field coordinators, and other staff.

Pilot testing logistics are as similar as possible to those anticipated for the full data collection, although not all subtasks may be tested and overall sampling considerations (such as regions, districts, schools, pupils per grade) will likely vary.

---

<sup>16</sup> If multiple versions of an instrument will be needed for baseline/endline studies, for example, preparing and piloting parallel forms at this stage helps determine and has the potential to lessen the need for equating the data after full collection; refer to Section 4.4 for guidelines on creating equivalent instruments and Section 8.5 for guidelines on statistical equating.

**Exhibit 19** outlines the key differences between the pilot test and the full data collection.

### Exhibit 19. Differences between EGRA pilot test and full data collection

	Pilot test	Full data collection
<b>Purpose:</b>	To test the reliability, validity, and readiness of instrument(s) and give assessors additional practice	To complete full assessment of sampled schools and pupils
<b>Timing:</b>	Takes place after adaptation	Considers the time of year in relation to academic calendar or seasonal considerations (holidays, weather); also factors in post-pilot adjustments and instrument revisions
<b>Sample:</b>	Convenience sample based on target population for full data collection	Based on target population (grade, language, region, etc.)
<b>Data:</b>	Analyzed to revise instrument(s) as needed	Backed up throughout the data collection process (e.g., uploaded to an external database) and analyzed after all data are collected
<b>Instrument revisions:</b>	Can be made based on data analysis, with limited re-piloting after the changes	No revisions are made to the instrument during data collection

#### 7.1.1 Pilot Study Data and Sample Requirements

To ensure that the pilot data are sufficient for the psychometric analysis conducted to establish test validity and reliability, it is required to collect a minimum of 150 non-missing and nonzero scores, and these zero scores must be of a reasonable range and comparable to the non-zero scores anticipated in the full study. Although ideally the pilot sample of schools and pupils would be selected randomly, most typically, the pilot sample is obtained through a convenience sample (see glossary). The reason for this is three-fold. First, the main purpose of the pilot is to ensure that the instrument is functioning properly; second, the pilot data are not used to draw any conclusions regarding overall student performance within a country, meaning that the sample does not need to be representative; and third, data collection using a convenience sample can be done more quickly and less expensively than collecting data by random sampling.

The students and schools selected for the pilot sample should be similar to the target population of the full study. However, to minimize the number of zero scores obtained within the pilot results, assessors may intentionally select higher-performing students or the planners may specifically target and oversample from higher-performing schools. In countries where the majority (70–80%) of primary students get zero scores, a very large randomly selected pilot sample would be needed to obtain 150 non-zero scores. For example, if it is anticipated that only 20% of cases would result in non-zero scores, a pilot sample of 750 students would be required to obtain the 150 non-zero scores needed for psychometric analysis. However, oversampling higher-performing schools could reduce the pilot sample size significantly.

To see how the EGRA instrument functions when administered to a diverse group of students, pilot data obtained through convenience sampling should include pupils from low-performing, medium-performing, and higher-performing schools. Note that if

school performance data are not available, it is advised to review socio-economic information for the specific geographic areas and use this information as a proxy for school performance levels. In general, it is not recommended that the convenience sample includes higher grades than the target population (e.g., fifth grade instead of second grade) as these students will have been exposed to different learning materials than target grade students and the range of non-zero scores may be quite different. However, in some contexts it is not possible to locate sufficient numbers of higher performing schools. This was the case in Zambia for some languages during the 2014 pilot test. In this case, it is permissible to go to higher grades in a pilot school as long as the target grade is also assessed.

Finally, the pilot sample, unlike the full study EGRA sample that limits the number of students per grade and per school to 10-12 pupils, tends to sample larger numbers of pupils per school. This type of oversampling at a given school allows for the collection of sample data more quickly and with a smaller number of assessors. Again, this is an acceptable practice because the resulting data are not used to extrapolate to overall performance levels in a country.

### 7.1.2 Establishing Test Validity and Reliability

**Test reliability.** Reliability is defined as the overall consistency of measure. For example, this could pertain to the degree to which EGRA scores are consistent over time or across groups of students. An analogy from everyday life is a weighing scale. If a bag of rice is placed on a scale five times, and it reads “20 kg” each time, then the scale produces reliable results. If, however, the scale gives a different number (e.g., 19, 20, 18, 22, 16) each time the bag is placed on it, then it is unreliable.

**Test validity.** Validity pertains to the correctness of measures and ultimately to the appropriateness of inferences or decisions based on the test results. Again, using the example of weighing scale, if a bag of rice that weighs 30 kg is placed on the scale five times and each time it reads “30,” then the scale is producing results that not only are reliable, but also are valid. If the scale consistently reads “20” every time the 30-kg bag is placed on it, then it is producing results that are reliable (because they are consistent) but invalid.

The most widely used measure of test-score reliability is **Cronbach’s alpha**, which is a measure of the internal consistency of a test (statistical packages such as SAS, SPSS, and Stata can readily compute this coefficient). If applied to individual items within the subtasks, however, Cronbach’s alpha may not be the most appropriate measure of the reliability of those subtasks. This is because portions of the EGRA instrument are timed. Timed or time-limited measures for which students have to progress linearly over the items affect the computation of the alpha coefficient in a way that makes it an inflated estimate of test score reliability; however, the degree to which the scores are inflated is unknown. Therefore, Cronbach’s alpha and similar measures are not used to assess the reliability of EGRA *subtasks individually*. For instance, it would be improper to calculate the Cronbach’s alpha for, say, the nonword reading subtask in an EGRA by considering each nonword as an item. On the other hand, using summary scores (e.g., percent correct, or fluency) of subtasks,

and calculating the overall alpha of an EGRA (across all subtasks) using those numbers, is necessary.<sup>17</sup>

For Cronbach's alpha or other measures of reliability, the higher the alpha coefficient or the simple correlation, the less susceptible the EGRA scores are to random daily changes in the condition of the test takers or of the testing environment. As such, a value of 0.7 or greater is seen as acceptable, although most EGRA applications tend to have alpha scores of 0.8 or higher.

In addition to the basic measures of reliability discussed above, it is useful to examine whether or not the assessment is unidimensional (i.e., it measures a single construct, such as early grade reading ability). One approach for measuring unidimensionality is to conduct exploratory factor analysis (EFA). This type of analysis hypothesizes an underlying (latent) structure in the data in order to identify the total number of constructs. Associated eigenvalues can be used to determine whether or not the first factor accounts for enough variance in order for the overall test to be considered unidimensional—that is, for the test to be testing a single overall construct that could be called “early grade reading.” While there is no specific cutoff for eigenvalues, scree plots are a visual representation used to determine whether or not there are multiple constructs (such that there is a natural break after the first factor, with a plateau of diminished values). Most statistical packages contain procedures for EFA. As with other measures, the analysis is done only on summary measures of the subtasks (e.g., percent correct, fluency) and on EGRA as a whole, not on the correctness of individual items within the subtasks. Most EGRA applications have a first factor explaining enough variance to suggest that the assessment is indeed assessing a single important overall construct.

Another aspect of reliability is measuring the consistency among raters to agree with one another (known as interrater reliability, IRR) during the field data collection process. If two assessors are listening to the same child read a list of words from the EGRA test, are they likely to record the same number of words as correctly read? This type of reliability measure involves having assessors administer a survey in pairs, with one assessor administering the assessment and one simply listening and scoring independently. Further explanation of how to administer IRR can be found in Section 6. Measuring the agreement between raters can be then be calculated by estimating Cohen's *kappa* coefficient (see glossary). This statistic (which takes a guessing parameter into account) is considered an improvement over percent agreement among raters, but both measures should be reported. While there is an on-going debate regarding meaningful cutoffs for Cohen's *kappa*, information on benchmarks for assessor agreement and commonly cited scales for *kappa* statistics can be found in Annex C, Section C.4.

In order to ascertain construct validity, item-level statistics should be produced to ensure that all items are performing as expected. Rasch analyses (which rely on an assumption of unidimensionality) provide construct validity information in several ways. First, the Rasch model places items and students on the same scale of measurement, in order, from easy (low ability for students) to difficult (high ability). Therefore, the order of the items from least to most difficult is the operational

---

<sup>17</sup> It should be noted that these measures are calculated on pilot data first, in order to ensure that the instrument is reliable prior to full administration; but they are recalculated on the operational (i.e., full survey) data to ensure that there is still high reliability.

definition of the construct. If this definition matches the intended design, there is an indication of construct validity. However, if there are instances where students do not have representative items accurately assessing their ability, it is said that there is underrepresentation of the construct. Finally, Rasch analyses assess item performance through fit statistics. If the items are not accurately measuring ability, or are producing “noise,” then they will have higher statistics ( $\geq 2.0$ ) indicating misfit and will need to be reevaluated. Assessments with many misfitting items are said to have *construct irrelevant variance*, which is also a detriment to construct validity. The outputs from a Rasch model can help test developers determine whether or not items behave as expected, and which items (if any) should be removed or revised due to poor fit. It is essential that these analyses be conducted on both pilot data (for initial test operational data) and full study data (to determine whether or not any specific items should be removed from scoring).

During the interval between the pilot test and the full data collection, statisticians and psychometricians analyze the data and propose any needed adjustments; language specialists and translators make corrections; electronic versions of the instruments are updated and reloaded onto all tablets; any hardware issues are resolved; and the assessors and supervisors are retrained on the changes.

### 7.1.3 Considerations Regarding the Timing of the Pilot Test

This section discusses the pros and cons of two options for the timing of the pilot test in relation to the timing of the assessor training and the full data collection.

The pilot testing of the instruments can take place before or after assessor training. There are advantages and disadvantages to both approaches, and the decision often comes down to logistics and context.

If no experienced assessors are available (from a prior administration of the assessment), it may be best to schedule the pilot test to take place immediately after the assessor training workshop ends. Typically pilot testing will take only one or two days to complete if all trained assessors are dispatched. An advantage of this approach is that the pilot test, in addition to generating important data about the instruments themselves, also provides valuable insight into the performance of the assessors. Those analyzing the pilot data can look for indications that assessors are making certain common mistakes, such as rushing the child or allowing more than the allotted time to perform certain tasks.

A disadvantage of pilot testing after assessor training is that the instruments used during assessor training are not yet finalized because they have not been pilot tested. In many cases, earlier less-formal pretesting of the instruments will have contributed to their being fine-tuned, such that the formal pilot test typically does not give rise to major instrument revisions. Still, in this scenario, assessors should be informed that the instruments they are practicing with during training may have some slight changes during later data collection. The implementer should thoroughly communicate any changes that take place after the pilot test to all assessors before they go into the field.

When pilot testing takes place immediately after assessor training, it is recommended that a period of at least two weeks elapse between the pilot test and full data collection, to allow for analysis of pilot

data, instrument revisions, printing, updating of electronic data collection interfaces, and distribution of materials to assessment teams.

In other cases, it is preferable to conduct pilot testing prior to assessor training. In contexts where an EGRA has taken place previously in the recent past (no more than two years prior), and hence trained assessors are available, a brief refresher training over one or two days can be sufficient to prepare for the pilot test. An advantage of this approach is that the instruments can be finalized (based on data analysis from the pilot test) before assessor training begins. Similar to the recommendation above, it is prudent to allow for at least two weeks between pilot testing and assessor training, so that all materials can be prepared not only for training, but also for data collection. In this scenario, data collection can begin as soon as possible after training ends.

\*The highlighted portion of this subsection comes directly from Kochetkova and Dubeck (In press). © UNESCO Institute of Statistics. Used by permission. All rights reserved.

## 7.2 Field Data Collection Procedures for the Full Studies

**Transport.** Each team will have a vehicle to transport materials and arrive at the sampled schools before the start of the school day.

**Assessment workload.** Experience to date has shown that application of the EGRA requires about 15 to 20 minutes per child. During the full data collection, this means that a team of three assessors can complete about nine or ten instruments per hour, or about 30 children in three uninterrupted hours.

**Quality control.** It is important to ensure the quality of instruments being used and the data being collected. Implementers must follow general research best practices:

- Ensure the safety and well-being of the children being tested, including obtaining children's assent.
- Maintain the integrity of the instruments (i.e., avoid public release).
- Ensure that data are collected, managed, and reported responsibly (quality, confidentiality, and anonymity<sup>18</sup>).
- Rigorously follow the research design.

**Equipment.** Properly equipping assessors and supervisors with supplies is another important aspect of both phases of the field data collection.

For data collection, the supplies needed include:

- Tablet, fully charged and loaded with current version of the instrument

---

<sup>18</sup> Anonymity: The reputation of EGRA and similar instruments relies on teacher consent/student assent and guarantee of anonymity. If data—even pilot data—were to be misused (e.g., schools were identified and penalized), this could undermine the entire approach to assessment for decision making in a given country or region.

- A laminated book of student stimuli, one per assessor (the same laminated book will be used for each student that the assessor tests)<sup>19</sup>
- Stopwatches or timers (in case tablets fail and backup paper instruments must be used)
- Pencils with erasers and clipboards
- Pencils or other small school materials to give to students in appreciation for their participation (if the planners have verified beforehand that doing so complies with any donor regulations)

**Supervision.** It is important to arrange for a supervisor to accompany each team of assessors. Supervisors provide important oversight for assessors and the collection process. Supervisors are also able to manage relationships with the school staff; accompany students to and from the testing location; replenish assessors' supplies; communicate with the support team; and fill in as an assessor if needed.

**Logistics.** Pilot testing is useful for testing the logistical arrangements and support planned for the data collection process. However, the full data collection involves additional aspects of the study that are sorted out before assessors leave for fieldwork: verifying sample schools, identifying locations, and arranging travel/accommodations to the schools. An itinerary also is critical and will always include a list of dates, schools, head teachers' contact numbers, and names of team members. This list is developed by someone familiar with the area. Additionally, the study's statistician will establish the statistical sampling criteria and protocols for replacing schools, teachers, and/or students, and the training team communicates them well to the assessors. Finally, for the full data collection phase, the planners organize and arrange the delivery of the assessment materials and equipment such as backup copies of instruments, tablets, and school authorization letters.

**Before departing for the schools,** assessors and supervisors:

- Double-check all materials
- Discuss test administration procedures and strategies for making students feel at ease
- Verify that all administrators are comfortable using a stopwatch or their own watches in case tablets fail.

**Upon arrival at the school,** the supervisor introduces the team of assessors to the school principal. In most countries, a signed letter from the government will be required to conduct the exercise; the supervisor also orally explains the purpose and objectives of the assessment, and thanks the school principal for the school's participation in the early grade reading assessment. The supervisor must *emphasize* to the principal that the purpose of this visit is **not** to evaluate the school, the principal, or the teachers; and that all information will remain anonymous.

The supervisor must ask the principal if there is an available classroom, teacher room, or quiet place for each of the administrators to conduct the individual assessments. Assessors proceed to whatever space is indicated and set up two

---

<sup>19</sup> Because the student stimulus sheets will be used with multiple students, lamination, while not completely necessary, does prolong the life of the student response forms (plastic page-protector sheets inserted into binders are also useful).

chairs or desks, one for the student and one for the assessor. It is also helpful to ask if there is someone at the school who can help throughout the day; this person also stays with the selected pupils in the space provided.

**During the first assessment each day**, the supervisor arranges for assessors to work in pairs to simultaneously administer the EGRA to the first student selected, with one actively administering and the other silently observing and marking. This dual assessment—which helps assure the quality of the data by measuring interrater reliability on an ongoing basis—is described further in Section 6.7.

**During the school day**, the primary focus is the students involved in the study. Assessors will have been trained on building rapport, but often the pilot is the first time they will have worked with children. Supervisors will be watching closely to make sure none of the children seem stressed or unhappy and that assessors are taking time to establish rapport before asking for the students' assent. Any key points from the observations of assessors working with the children are shared during the pilot debrief so that once teams go into the field, they are more adept at working with the pupils. Something as simple as making sure assessors silence their mobile phones makes a difference for students.

The supervisor must remind assessors that if students do not provide their assent to be tested, they will be kindly dismissed and a replacement selected using the established protocol.

If the principal does not designate a space for the activity, the assessment team will collaborate to locate a quiet space (appropriate for adult/child interaction) that will work for the assessment. The space should:

- Have sufficient light for reading and for the assessors to view the tablets
- Have desks arranged such that the students are not able to look out a window or door, or face other pupils
- Have desks that are clear of all papers and materials (assessors materials are on a separate table or on a bench so they do not distract the child)
- Be out of range of the selected pupils; students who are waiting are not be able to hear or see the testing.

## 7.3 Selecting Students

This section introduces two options for student sampling once assessors reach a sampled school. The first is enrollment based and the second is called interval sampling.

### 7.3.1 Student Sampling Option 1: Random Number Table

If recent and accurate data on student enrollment by school, grade, and class are available at the central level before the assessment teams arrive at the schools, a random number table can be used to generate the student sample. Generating such a random number table can be statistically more accurate than interval sampling. As this situation is highly unlikely in most country contexts, Option 2 is more commonly used.

### 7.3.2 Student Sampling Option 2: Interval Sampling

This sampling method involves establishing a separate sample for each grade being assessed at a school. The idea is to identify a sampling interval to randomly select students, beginning with the number of students present on the day of the assessment. This method requires three distinct steps.

**Step 1: Establish from the research design what group(s) will form the basis for sampling**

It is important to note that Step 1 must be finalized well before the assessors arrive at a school. This determination is made during the initial planning phases of research and sample design. During the assessor training, the assessor candidates will be instructed to practice the sampling methodology based on the research design.

The purpose of Step 1 is to determine the role of teacher data, the grade(s) and/or class(es) required, and expectations for reporting results separately for boys and girls. **Exhibit 20** presents the considerations required.

<b>Exhibit 20. Determinants of the sampling groups</b>			
<b>Research design— teacher data:</b>	The survey does not involve teacher data which will be linked to students	The survey involves teacher data for a single teacher in each grade which will be linked to student performance data	The survey involves teacher data for multiple teachers in each grade which will be linked to student performance data
<b>Basis for sampling— grade or class:</b>	Grade level	Class level – one class per grade	Class level – more than one class per grade
<p>Notes:</p> <ul style="list-style-type: none"> <li>• Surveys may involve one or more grades.</li> <li>• In addition to selection by grade/class, the research design may specify that the students are be selected by sex (see next row).</li> <li>• Assessors' school materials include a set of dice for randomly selecting a class or classes, should there be multiple teachers for the sampled grade. The sampling protocol specifies how the dice are to be used.</li> </ul>			
<b>Group(s) from which the sample(s) must be selected:</b>	<p>Either:</p> <ul style="list-style-type: none"> <li>• All the students in each grade (irrespective of gender)</li> </ul> <p>Or:</p> <ul style="list-style-type: none"> <li>• All the male students in each grade, and</li> <li>• All the female students in each grade</li> </ul>	<p>Either:</p> <ul style="list-style-type: none"> <li>• All the students from each selected class in each grade (irrespective of gender)</li> </ul> <p>Or:</p> <ul style="list-style-type: none"> <li>• All the male students in each selected class, and</li> <li>• All the female students in each selected class</li> </ul>	

### Step 2: Determine the number of students to be selected from each group: $n$

The second step consists of making calculations based on the total number of students to be sampled per school and the number of groups involved.<sup>20</sup>

Illustration: If the total number of students to be sampled is 20 per school and the students are to be selected from one class in each of two grades (e.g., grades 2 and 3) according to sex, then there are four groups and five students ( $20 \div 4$ ) that are to be selected from each group, as follows:

1. 5 male students from the selected class in grade 2
2. 5 female students from the selected class in grade 2
3. 5 male students from the selected class in grade 3
4. 5 female students from the selected class in grade 3

### Step 3: Randomly select $n$ students from each group

The purpose of this step is to select the specific children to be assessed. The recommended procedure is:

1. Have the children form a straight line outside the classroom.
  - o If assessing children from more than one grade, begin with the children from the lower grade at the start of the day.
2. Count the number of children in the line:  $m$ .
3. Divide  $m$  by  $n$  (from Step 2) and round the answer to the nearest whole number:  $p$ .
4. Starting at one end of the line, randomly select any child from the first  $p$  children and then count off and select each  $p^{\text{th}}$  child after that.

Illustration: To select  $n = 8$  children from a given group:

1. There are 54 children in the line ( $n = 54$ )
2. Calculate  $p$ :  $54 \div 8 = 6.75$ ; round:  $p = 7$
3. Randomly select a child from the first  $p = 7$  children<sup>21</sup> – for example, child number 3
4. Select every  $p^{\text{th}}$  child starting with child 3:  
3; 10; 17; 24; 31; 38; 45; 52

Note that this procedure should result in 9 selected children—the 9th child is an alternate in case one child does not want to participate. In the above example that has 54 children, the assessor should continue counting and selecting every 7th child until the end of the line, and then circle back to the beginning of the line to select the next 7th child (which would be the 5th child from the start of the line).

<sup>20</sup> See Annex B and Section 7 for more information on sample design.

<sup>21</sup> This process is known as “random start.”

Once the assessors have administered the EGRA to all the students in the first group (as designated in Step 2), the assessment team repeats Step 3 to select the children from the second group. The supervisor ensures the assessors always have a student to assess so as not to lose time during the administration.

#### **7.4 End of the Assessment Day: Wrapping Up**

To the extent possible, all interviews at a single school are completed within the school day. A contingency plan must be put in place at the beginning of the day, however, and discussed in advance with assessors and supervisors as to the most appropriate practice given local conditions. If the school has only one shift and some assessments have not been completed before the end of the shift, the supervisor will find the remaining students and ask them to wait beyond the close of the school day. In this case, the school director or teachers make provisions to notify parents that some children will be late coming home.

#### **7.5 Uploading Data Collected in the Field**

Assuming data are collected electronically (this is current recommended best practice—see Section 5), the planners arrange the means for assessors to send data to a central server every day to avoid potential data loss (i.e., if a mobile device is lost or broken). If this is not possible, then backup procedures are in place. Procedures for ensuring data are properly uploaded or backed up will be the same during both pilot testing and full data collection. The pilot test is an important opportunity to make sure that these procedures function correctly.

Assessors will send their data to the central server using wireless Internet, either by connecting to a wireless network in a public place or Internet café, or by using mobile data (3G). When planning data collection, planners must consider factors such as available carrier network, compatibility between wireless routers and modems, and technical capacity of evaluators, and seek the most practical and reliable solutions. During the piloting, evaluators practice uploading and backing up data using the selected method. A data analyst verifies that the data are actually uploading to the server and then reviews the database for any technical errors (i.e., overlapping variable names) before the full data collection proceeds.

## BENEFITS OF REGULARLY UPLOADING AND REVIEWING DATA

During data collection, regular data uploading and review can help catch any errors before the end of data collection, saving projects from sending data collectors back into the field after weeks of data collection. Additionally, daily uploads can help prevent loss of large amounts of data if a tablet is lost, is stolen, or breaks. Data can be checked to ensure that the correct grade is being evaluated, that assessors are going to the sampled schools, and that the correct numbers of students are being assessed, as well as to verify any other inconsistencies. Constant communication and updates to let the project team know when data collection is proceeding, when the data analysts sees uploaded data, and if there are any delays or reasons that would prevent the uploading of data on a daily basis can help in reviewing the data as well as in knowing what results to expect and when.

Backup procedures for electronic data collection include having paper versions of the instrument available for the data collectors' use. After every assessment completed in paper form, the supervisor reviews the paper form for legibility and completeness (i.e., no missing school code or ambiguous tick marks). The supervisor or designated individual is in charge of keeping the completed forms organized and safe from loss or damage, and ensuring access only by authorized individuals.

# 8 PREPARATION OF EGRA DATA

This section covers the process of cleaning and preparing EGRA data. Once data are collected, recoding and formulas need to be applied to create summary and super-summary variables. Note that this section assumes that weights and adjustments to sampling errors from the survey design have been appropriately applied.

Nearly all EGRA surveys consist of some form of a stratified complex, multistage sample. Great care is required to properly monitor, check, edit, merge, and process the data for finalization and analysis. These processes must be conducted by no more than two (extremely experienced) statisticians. One person conducts these steps while the other person checks the work. Once the data are processed and finalized, then anyone with experience exploring complex samples and hierarchical data can familiarize themselves with the objectives of the research, the questionnaires/assessments, the sample methodology, and the data structure, and then easily analyze the data.

This section assumes the statistician(s) *processing* the data has extensive experience in manipulating complex samples and hierarchical data structures, and gives some specifics of EGRA data processing.

## 8.1 Data Cleaning

Cleaning collected data is an important step before data analysis. To reiterate, data cleaning and monitoring must be conducted by a statistician experienced in this type of data processing.

Data quality monitoring is done as data are being collected. Using the data collection schedule and reports from the field team, the statistician is able to match the data that are uploaded to the expected numbers of assessments for each school, language, region, or other sampling unit. During this time, the statistician responsible for monitoring will be able to communicate with the personnel in the field to correct any mistakes that have been made during data entry, and to ensure the appropriate numbers of assessments are being carried out in the correct schools and on the assigned days. Triangulation of the identifying information is an important aspect of confirming a large enough sample size for the purposes of the study. Being able to quickly identify and correct any of these inconsistencies will aid data cleaning, but will also ensure that data collection does not have to be delayed or repeated because of minor errors.

**Exhibit 21** is a short checklist for statisticians to follow during the cleaning process, to ensure that all EGRA data are cleaned completely and uniformly for purposes of the data analysis.

## Exhibit 21. Data cleaning checklist

**Review incomplete assessments.**

Incomplete assessments are checked to determine level of completeness and appropriateness to remain in the final data. Each project will have agreed criteria to make these decisions. For example, assessments that have not been fully completed could be kept if it is necessary for purposes of the sample size to use incomplete information; or the assessment being used can be verified as accurate and is not lacking any important identifying information.

**Remove any “test” assessments that were completed before official data collection began.**

Verify that all assessments included in the “Cleaned” version of the data used for analysis are real and happened during official data collection.

**Ensure that all assessments are linked with the appropriate school information for identification.**

Remove any assessments that are not appropriately identified, or work with the field team to ensure that any unlabeled assessments are identified accurately and appropriately labeled.

**Ensure child’s assent was both given and recorded for each observation.**

Immediately remove any assessments that might have been performed without the assessor having asked for or recorded the child’s expressed assent to be assessed.

**Calculate all timed and untimed subtask scores.**

Information on scoring timed and untimed subtasks can be found in Section 8.2.

**Ensure that all timed subtask scores fall within an acceptable and realistic range of scores.**

During data collection, assessors may make mistakes, or data collection software malfunctions may lead to extreme outliers among the scores. Investigate any exceptionally high scores and verify that they are realistic for the pupil being assessed (based on the child’s performance in other subtasks), and were not caused by some error. Remove any extreme observations that are determined to be errors in assessment, so as not to skew any data analysis. It is not necessary to remove all observations from that particular pupil, as this would affect the sample size for analysis in other subtasks. Simply remove any scoring from the particular subtask that is shown to be in error.

## 8.2 Processing of EGRA Subtasks

This section begins with the nomenclature for the common EGRA subtasks and variables, then discusses what information must be collected during the assessment and how to derive the rest of the needed variables from the raw variables collected. Note that **Annex D** of the toolkit is an example of a codebook for the variables in an EGRA data set.

Basically, the EGRA variable names have the structure:

**<prefix>\_<core><suffix>**

Examples:

**e\_letter\_sound1**  
**e\_letter\_sound2**  
**e\_letter\_sound\_time\_remain**

To maintain consistency within and across EGRA surveys, it is important to label subtask variables with the same names. **Exhibit 22** provides a list of variable names for EGRA subtasks as well as the names for variable timed scores (if the subtask is timed).

### Exhibit 22. EGRA subtask variable nomenclature and names of the timed score variables

Name of subtask variable	Name of subtask	Name of subtask timed variable	Label for subtask timed
letter_sound	Letter Identification (Sounds)	clspm	Correct Letter Sounds per Minute
invent_word	Nonword Reading	cnonwpm	Correct Nonwords per Minute
oral_read	Oral Reading Fluency	orf	Oral Reading Fluency
read_comp	Reading Comprehension		
list_comp	Listening Comprehension		
oral_vocab	Oral Vocabulary		

#### 8.2.1 <prefix>\_

If a student was assessed in more than one language, it is important to distinguish the languages with a prefix. Secondary languages need a prefix, such as an e\_ for English or f\_ for French.

**Note about multiple passages:** In many pilot studies, there is more than one version of the same subtask. For example, there may be three different versions of the oral reading fluency passage as well as three different sets of comprehension questions. In these cases, the prefixes are the language letter and the number of the different subtask. So for English, the variable names would be e1\_oral\_read<suffix>, e2\_oral\_read<suffix>, e3\_oral\_read<suffix>, to help distinguish which reading passage the variable is referring to.

#### 8.2.2 <suffix>

The EGRA subtasks will result in data being collected for each item a student got right, got wrong, or did not attempt because time ran out. That is to say, for the letter identification (sounds) subtask, the data will have a variable for each item tested. From this information, it is possible to calculate all summary untimed score variables. The suffixes indicate the subtask item number and the score summary.

The suffix will be the item number in the subtask or any additional variables associated with this subtask (such as: \_auto\_stop, \_attempted, \_time\_remain). The suffix could be the item number found the subtask. For example, if there were five

items in the English reading comprehension section, the variable names would be e1\_read\_comp1, e1\_read\_comp2, e1\_read\_comp3, e1\_read\_comp4, e1\_read\_comp5, e1\_read\_comp\_attempted.

Please note, these item variable names do not have an underscore “\_” between the core and the suffix number 1–5. So, variables would NOT be: e\_read\_comp\_1, e\_read\_comp\_2, e\_read\_comp\_3, e\_read\_comp\_4, e\_read\_comp\_5. Non-item variables have an underscore “\_” between the core and the suffix. Non-item EGRA variables are named e\_read\_comp\_attempted and e\_read\_comp\_score.

**Exhibit 23** contains some examples of how the EGRA variables are named, based on the language and the number of sections repeated within the instrument.

### Exhibit 23. Suffix nomenclature for the item and score variables

Suffix	Variable suffix label	Possible values
1-#	Item #	0 "Incorrect" 1 "Correct" . <missing> "Not asked/didn't attempt"
_score	Raw Score	0 - # Items in Subtask
_attempted	Total Items Attempted	0 - # Items in Subtask
_score_pcmt	Percent Correct	0-100
_score_zero	Zero Score Indicator	0 "Score>0" 1 "Score=0"
_attempted_pcmt	Percent Correct of Attempted	0-100

The following summary variables are then calculated:

- **\_score**. Sum of the correct item responses (which are coded as 1).
- **\_attempted**. Count of the correct and incorrect item responses, which are coded as either 1 or 0.
- **\_score\_pcmt**. Subtask\_score divided by the number of possible items in subtask.
- **\_score\_zero**. Yes (recorded as 1) if the student scored zero; otherwise, No (coded as 0).
- **\_attempted\_pcmt**. \_score divided by \_attempted.

### 8.3 Timed Subtasks

A timed subtask in the EGRA instrument is designed to be calculated on a *per minute* rate. Responses, such as individual letters or words, must be coded as either *correct*, *incorrect*, or *no response/did not answer*. The field assessor must distinguish between *incorrect* (coded as zero) and *no response*, as it will not be possible to analyze items attempted of there is no differentiation.

In addition to the item responses, the following summary variables must be included in the raw data for timed subtasks:

1. **Subtask\_time\_remain.** This is the time remaining in a subtask if a student finished the task before the allotted time expired. This summary variable will be used to calculate the *per minute* rate. It is recorded in seconds. Typically, a timed subtask will have a maximum of 60 seconds to be completed. Thus, time remaining will be 60 seconds minus the time taken to complete the subtask.
2. **Subtask\_auto\_stop.** In order to move efficiently through the assessment and not have students pause for a lengthy period trying to answer questions they clearly do not know, the assessment is stopped after a student is unable to answer the first few items—typically the first 10 (or fewer) items. A student who cannot respond before the auto-stop receives a code of 1 for that subtask, with 1 meaning yes the student was auto-stopped. This score is for the overall subtask and not recorded at the item level.

In order to create summary variables, individual item responses are set to 1 for correct answers, 0 for incorrect answers, and *missing* for no response/did not answer.

The per-minute rate is often referred to as a fluency rate. The timed subtasks are usually administered over a 60-second timed period, such that only those students who finish responding to the items in a subtask or reading the passage before the time ends will have fluency value different from their raw score. The final unit of measurement is either correct letters or correct words per minute.

The per\_minute rate is calculated using the following formula:

$$\text{Subtask\_per\_minute} = \frac{\text{Subtask\_score}}{\text{Time given for subtask} - \text{subtask\_time\_remain}} \times 60$$

## 8.4 Untimed Subtasks

As with the timed subtasks, these item responses need to be coded as *correct*, *incorrect*, or *no response/did not answer*. In order to create summary variables, item responses are set to 1 for correct answers, 0 for incorrect answers, and *missing* for no response/did not answer.

### Note about the reading comprehension activity:

As is standard practice, if reading comprehension is calculated from the same passage from which oral reading was assessed, students have been assessed on the number of reading comprehension questions they answered in the section of the passage they were able to read.

For example, if five reading comprehension questions were based on having read the passage through the 9th, 17th, 28th, 42nd, and 55th words, respectively, and a student read to the 33rd word, then that student will be assessed on the first three reading comprehension questions. The attempted responses are marked: correct, incorrect, or no response. The two final questions will be coded as *not asked*.

Although this benchmark may vary by context, in general, students are considered to be able to read fluently, with comprehension, if they read an entire passage and can answer 80% or more of the reading comprehension questions correctly. To calculate this, a new summary variable is created: **read\_comp\_score\_pcmt80**, which is correct (coded to 1) if the reading comprehension score percent is 80% or higher; otherwise it is set to incorrect (coded as 0).

## 8.5 Statistical Equating

Equating is a statistical procedure used to convert scores from multiple forms of a test to the same common measurement scale. This conversion process adjusts for any difficulty with differences between forms, so that a score on one form can be matched to its equivalent value on another form. As a result, equating makes it possible to estimate the score that a child being assessed with one form would have received had they been assessed with a different test form (Kolen & Brennan, 2004; Holland & Dorans, 2006).

Research on small-sample statistical equating (which is appropriate for nearly all EGRA equating) has shown that when true score differences between subtasks on two test forms are less than approximately 1/10 of a standard deviation, equating error can actually exceed the bias of not equating (Hanson, Zeng, & Colton, 1994; Skaggs, 2005). Therefore, equating is not recommended for small samples when the difference in scores across forms is no greater than 1/10 of a standard deviation.

When equating is necessary, there are a few important considerations to keep in mind.

The first point is that instrument developers must consider and recognize subtasks' suitability for equating. Four technical terms that underlie this discussion are *common-item equating*, *common-person equating*, *classical test theory (CTT) equating*, and *item response theory (IRT) equating*.

**Common-item equating:** It is used when instruments or subtasks are designed with some items that are common to all test forms. These common items (also known as *anchor items*) account for at least 20% to 25% of the total items on the assessment, and they represent a mini-version of the overall assessment (in terms of difficulty and variation). It is also important to ensure that anchor items retain their placement across test forms (e.g., if a particular anchor item is the fifth item on test form A, it is also the fifth item on test form B). The remaining items (i.e., non-anchor items) can be either reshuffled items from the original instrument or entirely new items.

The basic principle behind common-item equating is that the difficulty of anchor items is identical across assessment forms. Therefore, scores are adjusted to account for overall test difficulty based on the subscore for the anchor items. There are many methods for conducting common-item equating (including chained equating and post-stratification), but the breadth and depth of information needed to cover these topics are outside the scope of this toolkit.

Ultimately, common-item equating is best for subtasks that have sufficient items (i.e., a recommended minimum of 20–25 items), because of the reduced likelihood of statistical error (assuming a similarly small sample size).

**Common-persons equating:** Also known as a single group design or randomly equivalent group design, this method is used when instruments or subtasks are designed to measure identical constructs but do not contain anchor items. This is currently the most common type of equating conducted for EGRA because it does not require knowledge of equating procedures at the instrument design stage. For this approach, multiple forms of the EGRA are piloted with a sample of students (each of whom take all forms). The basic principle is that differences in test scores across forms of the assessment can be seen as differences in test difficulty (as opposed to student ability), since the same students are taking each form. This approach is necessary for the oral reading fluency passage of EGRA since it is not possible to create anchor items for that subtask (and since item-level information is not relevant—which is a prerequisite for IRT equating, as discussed below).

## REQUISITE STEPS FOR COMMON-PERSONS EQUATING DURING PILOT

In order to maximize efficiency and to take fullest advantage of the common-persons equating design, the following scenario should be used during the pilot stage where there is sufficient time (and foresight) to create a large number of parallel passages and sufficient funding to conduct a pilot with at least 500 students.<sup>22</sup>

In this scenario, it is suggested that EGRA developers create 10 reading comprehension passages with five questions on each (10 sets), using expert judgment in their construction to make them as parallel as possible on the front end. Each sample of students would then be administered three separate passages (and accompanying comprehension questions). The design could (hypothetically) look as shown in **Exhibit 24** (with 10 forms of 3 sets and 15 questions, each).

**Exhibit 24. Sample counterbalanced design**

Number of students	First block	Second block	Third block	Pilot test form
50	1	2	4	A
50	2	3	5	B
50	3	4	6	C
50	4	5	7	D
<b>50</b>	5	6	8	E
<b>50</b>	6	7	9	F
<b>50</b>	7	8	10	G
<b>50</b>	8	9	1	H
<b>50</b>	9	10	2	I
<b>50</b>	10	1	3	J
500				

In this design, every passage appears in each block (first, second, third), and each passage appears with six other passages. Passage order is rotated in order to minimize order effects. This approach requires a sample of 500 students (randomly

<sup>22</sup> This singular pilot could take the place of multiple pilots of 150–200 students (which is not uncommon in development work). It is simply a matter of cost-benefit and the value of having 10 evaluated passages.

assigned into 10 subsamples, with each receiving one of the 10 test forms). Therefore, it is possible to obtain robust measures of the relative difficulty of each item and set. Sets are then matched in order to obtain maximum comparability for pre- and post-testing, with confidence that changes in scores at the sample level would be meaningful.

**Classical test theory (CTT) equating:** Equating models based on CTT establish relationships between total scores on different test forms. This is a more “traditional” approach to test equating, and it is the most common approach for equating with small samples. CTT equating approaches include mean, linear, circle-arc, and equipercenile equating. This toolkit does not provide in-depth explanations of each approach.

CTT equating is beneficial for linear data and for use with small samples. CTT equating is not recommended for subtasks with relatively few items (e.g., fewer than 10). For subtasks with 10–25 items, it may be possible to use a CTT pre-equating approach by piloting multiple, newly developed test forms along with baseline forms and comparing item-level statistics across forms. Ultimately, however, this approach is most useful for equating oral reading fluency.

**Item response theory (IRT) equating:** IRT equating is based on the principle of establishing equating relationships through models that connect observable and latent variables. This approach has the advantage of using the same mathematical model characteristics of people and characteristics of instruments. IRT equating also has the advantage of being more compatible with the nature of testing while providing opportunities to equate subtasks with few items. However, IRT equating is procedurally and conceptually complex and requires significantly larger samples than CTT equating (with the exception of the Rasch model, which requires the same sample size as CTT—which is approximately 100–150 participants).

Therefore, IRT equating is extremely useful for post-equating (i.e., equating on operational or full survey data—as compared with pre-equating, which is conducted using pilot data), when sufficient technical expertise and capacity are available. In the majority of EGRA work, IRT equating will ultimately be beneficial for pre-equating on subtasks that have few items as well as useful item-level data. Such subtasks include reading comprehension, listening comprehension, dictation, vocabulary, and maze.

# 9 DATA ANALYSIS AND REPORTING

This section of the toolkit provides a brief overview of the types of data analyses that correspond to various research designs, as well as required components to be included in EGRA reports.

When analyzing EGRA data, researchers must use descriptive and/or inferential statistics to describe the data, examine patterns, and draw conclusions. However, it is important to understand the differences between these two types of statistics, as well as the purpose and value of each.

## 9.1 Descriptive Statistics (Non-inferential)

Descriptive (or non-inferential) statistics are used to describe and summarize data—often in an effort to see what patterns may emerge. Descriptive statistics do not allow for conclusions to be drawn beyond the data, nor is it possible to test research hypotheses. The main purpose for descriptive analysis is to present data in a meaningful way that allows for ease of interpretation (as opposed to simply presenting raw data). The most common measures reported in descriptive analyses are frequencies, measures of central tendency (e.g., means and medians) and measures of spread (e.g., standard deviations and summary ranges).

Also, as the name implies, descriptive statistics are used only to describe sample data. In much EGRA work, samples are selected to be representative of larger populations. In these cases, reported frequencies, means, etc., are based on weighted data and thus effectively become inferential statistics. Therefore, descriptive statistics are to be reported only for studies that are designed to draw no conclusions beyond samples; or as unweighted frequencies, unweighted means, etc., for complex survey data.

Lastly, with non-inferential statistics, it is essential that the sample be fully described according to the level of disaggregation to be analyzed and reported. For example, if pupil scores in the report are going to be disaggregated by language and grade, then the sample descriptive statistics include these levels of disaggregation.

Examples of useful descriptive statistics in EGRA reporting would be frequencies and means of basic demographic characteristics of the sample, as well as unweighted means across subtasks for all levels of disaggregation.

## 9.2 Types of Regression Analysis

Given that regression is the most common way to analyze the relationships and predicted values of variables in EGRA data, it is important to briefly examine the different types of regression analyses that can be conducted. Ordinary least squares

(OLS) regression analysis works well for EGRA data that have normally distributed residual values, when a continuous variable such as the oral reading fluency score is being used.

However, many developing countries have test scores that cluster around zero, making the distribution of scores very uneven. When dealing with such data, evaluators should consider using binomial regression analysis, such as probit or logistic regression, which allows evaluators to examine binomial outcomes such as whether a student meets local benchmarks for reading ability or whether a student scores zero on a specific reading subtask.

### 9.3 Reporting Data Analysis

The purpose of analyzing EGRA data is both to improve program effectiveness and to provide findings to clients, partner organizations, and government officials via briefs and full program reports. Recognizing that different objectives as well as audiences for reporting will shape the structure and the content of those reports, the following guiding principles are necessary:

1. **Objectives and limitations.** The report must clearly state the objectives of the study and its limitations.
2. **Plain language.** The main findings must be presented in clear, concise, and nontechnical language.
3. **Data visualization.** Data visualization must be used to facilitate understanding of the findings by general audiences. Visualizations are “standalone,” such that the visual is interpretable without the audience needing to read extra text.
4. **Descriptive and inferential analyses.** The main report presents summary findings of descriptive data analysis, including mean distributions and grouped distributions. Inferential statistical analyses are used to design weights, post-stratification weights, and the standard errors to account for the complex survey design (if appropriate).
5. **Score distributions.** For every pupil score estimate reported, a visual of the score distribution must be graphically presented. This supports the reader’s interpretation of the estimate provided; for example, while the mean score can be produced, the accompanying distribution puts into perspective how “representative” the estimate is of pupil scores. This is especially important if the pupil scores are non-normal. In some cases, it may make sense to present median pupil scores in addition to the mean scores and distributions.
6. **Levels of disaggregation.** The results of data disaggregation by sex, grade, language, and other variables of interest must be described as appropriate to the research design.
7. **All results reported.** Whenever comparison-of-means statistical tests are conducted to compare across groups of subjects (such as sex or language), or bivariate/multivariate statistical analyses (e.g., correlations) are conducted to examine the relationship between different variables, results must be reported even if they are not statistically significant.

8. **Substantiation for inferential estimates.** The following must accompany all reported inferential estimates (including but not limited to means, median, mode and proportions):
  - Precision – either as 95% confidence interval for estimates, or a *t*-score and *p*-value for comparisons in addition to standard errors.
  - Sample size
9. **Effect sizes.** Whenever results of comparisons of data across groups are presented (such as differences between baseline and endline, or between boys and girls, or between rural school students and urban school students), effect size of the difference must be reported.
10. **Equivalence.** In experimental and quasi-experimental designs, equivalence of baselines must be established (What Works Clearinghouse, 2015).

# 10 USING RESULTS TO INFORM ACTION

## 10.1 Setting Country-Specific Benchmarks

One of the virtues of EGRA is that the science behind it corresponds fairly well to the average layperson’s concept of what it means to read: the notion of “knowing one’s letters,” being able to read unhesitatingly and at a reasonable rate, and being able to answer a few questions about what one has read. Thus, being able to report that children cannot recognize letters, or can read them only extremely slowly, is something that most individuals can interpret. Relying on the data produced by EGRA (or other types of individual, orally administered early grade assessments) is a sound way to tell the story of whether schools are serving students in the most basic way.

Nonetheless, for focusing the attention of policy makers and officials on the question of how students are learning to read, it is useful to be able to benchmark the results in some way. Benchmarks are particularly useful for reading, as they establish expectations and norms for reading performance. Benchmarks are needed to gauge progress in any given country or context. A sound benchmark can be used to easily translate a set goal into measures of progress at specific points in time. For example, if the goal is that all children will learn to read well by the end of grade 3, a benchmark can show the percentage of pupils achieving different levels of reading ability in a given grade and year—indicating whether progress is being made toward that overarching goal. Additionally, benchmarks are found to be helpful when they are used as a means to communicate publicly about improvement (e.g., school report cards or national-level monitoring and reporting).

Standards allow for a common and measurable expectation to be applied across state or national populations, but allowing decentralized decision-making about how to get children to achieve those goals. The same objective measurements also serve as a mechanism for accountability, holding schools—and sometimes teachers—responsible for educational achievement. Studies show that high-stakes assessment systems do affect teacher and administrator behavior, but not in consistent or predictable ways. Therefore, care must be taken when benchmarks are being developed to ensure that the education system can use them to measure progress and identify areas where additional effort is needed, rather than using them to mete out high-stakes consequences.

### 10.1.1 What Are Benchmarks?

Benchmarks have been defined as “a standard or point of reference against which things may be compared or assessed” (Oxford online dictionaries, <http://www.oxforddictionaries.com>); “A criterion for performance at a particular point (a milestone),” and “empirically derived, criterion-referenced target scores that

represent adequate reading progress” (Dynamic Measurement Group, Inc., 2010, p. 1).

For purposes of this toolkit, a “benchmark” is synonymous with a “standard” in that it defines a desired level of performance achievable at a particular point in time. Thus, a “benchmark assessment” is a diagnostic administered at regular intervals, used to evaluate whether students are progressing on track toward achieving desired standards. “Benchmark scores” may also be established at cut-points that help interpret the meaning of the specific score; for example, setting “basic,” “intermediate,” and “proficient” cut-points can help identify student profiles based on a definition of partial or total mastery.

Benchmarks may also be associated with “targets” (goals, objectives) that define expectations for the population; for example, if the benchmark determines how high to set the bar, the target defines how many children will clear that bar. For example: “60% of students meet the benchmark in Year 1; 80% of children meet the benchmark in Year 2.” Setting targets is particularly important where performance is low. The target defines an intermediate step toward achieving the goal.

In communication activities, messages are effective only if the desired audience can understand them. Providing EGRA results without a point of reference is usually ineffectual in environments where fluency measurements (i.e., 20 correct words per minute) are unfamiliar or assessments tend to be reported as a percentage of correct responses. A benchmark is a point of reference with which to interpret the performance because it provides an expected level of achievement. In the case of educational benchmarks, they add specificity to broad curricular goals such as “shall be able to read fluently” by stating instead, “shall be able to read at a rate of 40 correct words per minute by the end of grade 2.” However, those expectations need to be grounded in the country reality rather than adopted from other countries or languages. EGRA data can be used to define benchmarks, and subsequent administrations can generate data with which to evaluate performance over time according to those benchmarks.

### Definitions

- Goal is a long-term aspiration, maybe without numerical value  
**Goal: All our children should read**
- Metric is a valid, reliable unit of measurement  
**Metric: “correct words per minute in passage reading”**
- Benchmark is a numerical step towards the goal, using the metric  
**Benchmark: 45 correct words per minute, understand 80% of what they read**
- Target is a variable using the benchmark  
**Target: % of children at or above benchmark, or average achieved by the children, using the metric.**

Source: LaTowsky (2014)

## 10.1.2 Criteria for Establishing Benchmarks

Setting benchmarks can employ a process that combines statistical analysis of student data over time with additional information such as research about the way

children learn to read, experience elsewhere, insights from cognitive science and knowledge of local contexts. Benchmarks may change over time in line with improvements in student performance. There are many ways to develop standards or benchmarks, but the key criteria that good standards meet include:

- The benchmarks are ambitious, but realistic and achievable.
- They are not subject to score inflation (i.e., score increases do not generalize to other measures of the same content because they primarily reflect narrow test-preparation activities geared toward a specific test) (Hamilton, Stechter, & Yuan, 2008).
- Benchmarks must be able to identify students who are likely to fail at achieving an independent level of reading. Benchmarks are specific to a point in time (beginning of the year, end of the year, grade, etc.) and subsequent benchmarks are derived based on the probability that children meeting the first benchmark will also meet the next one (under current instructional conditions). (Dynamic Measurement Group, Inc., 2010)

*“There are no true or correct cut scores for a test, only more or less defensible ones. Defensibility is based in large measure on the method used to set standards. Second, there is no one best or correct method for setting standards but rather a range of approaches that may be more or less appropriate for a specific situation.”*

*– Ferrara, Perie, & Johnson, 2008*

- Benchmarks are based on research that examines the predictive validity of a score on a measure at a particular point in time, compared to later measures and external outcome assessments. If a student achieves a benchmark goal, then the odds are in favor of that student achieving later reading outcomes if he/she receives research-based instruction from a core classroom curriculum (Dynamic Measurement Group, Inc., 2010).
- The best kinds of data to use are the test scores of real test takers whose performance has been meaningfully judged by qualified judges (Zieky & Perie, 2006).
- Benchmarks are appropriately linked across the grades to avoid misclassification of students, or misleading reports to stakeholders. For example, while it may be appropriate to assign a higher cut-point to define an advanced student in grade 2 than defines a basic student in grade 3, the opposite is not true (Zieky & Perie, 2006).

All benchmarks are ultimately based on norms, or judgments of what a child should be able to do (Zieky & Perie, 2006). A country can set its own benchmarks by looking at performance in schools that are known to perform well, or that can be shown to perform well on an EGRA-type assessment, but do not possess any particular socioeconomic advantage or unsustainable level of resource use. Such schools will typically yield benchmarks that are reasonably demanding but that are demonstrably achievable even by children without great socioeconomic advantage or in schools without great resource advantages, as long as good instruction is taking place. The 2001 Progress in International Reading Literacy Study (PIRLS 2001), for example,

selected four cutoff points on the combined reading literacy scale labeled international benchmarks. These benchmarks were selected to correspond to the score points at or above which the lower quarter, median, upper quarter, and top 10 percent of fourth-graders in the international PIRLS 2001 sample performed (Institute of Education Sciences, n.d.).

### 10.1.3 A Process for Setting Benchmarks

The steps below explain the general process that has been used in at least 12 low-income countries, including Zambia, for setting benchmarks and targets.

**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 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).

**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.

As mentioned in Section 2.2, the Zambian Grade 2 National Assessment Survey conducted in 2014 was used to inform and draft national benchmarks and targets in July 2015. In the case of Zambia, during the benchmarking workshop, participants faced contextual challenges that resulted in important decisions which, in turn, informed the steps and process described above. The first decision made by the participants and experts attending the benchmarking workshop was to develop a single set of benchmarks for all the languages. While it is most commonly advised to develop a set of benchmarks and target for individual languages,<sup>23</sup> such an approach was judged not to be necessary in Zambia. The reason for this decision was that pupil performance across the languages in the Grade 2 NAS was more similar than not. Secondly, it was decided to set benchmarks for “emergent readers and mathematicians” as well as benchmarks for “readers and mathematicians.” In other words, benchmarks were set for two different groups of pupils based on their expected performance levels. To explain further, the baseline data had indicated that the number of pupils at the “readers and mathematicians” level was too low to reasonably expect noticeable changes within 5 years. Therefore, to make it feasible to detect any shift in scores over time, benchmarks and targets were set for a

---

<sup>23</sup> Developing a benchmark and target per language is often advised due to the differences in the orthographies and linguistic characteristics between any given languages.

somewhat less proficient group of “emergent readers and mathematicians.”

**Exhibit 25** below summarizes the benchmarks and targets that were set for reading and mathematics in Zambia during the July 2015 benchmarking workshop.

### Exhibit 25. National benchmarks and targets for reading and mathematics in Zambia

Benchmarks and Targets		Reading			Mathematics	
		Nonword Decoding	Oral Reading Fluency	Reading Comprehension	Missing Number	Addition and Subtraction Level 2
<b>Benchmarks</b>		<b>cwpm</b>	<b>cwpm</b>	<b>% correct</b>	<b>% correct</b>	<b>% correct</b>
	Emergent readers and mathematicians	15	20	40%	30%	40%
	Readers and mathematicians	30	45	80%	60%	70%
<b>Targets (percentages of pupils)</b>						
Zero score	Baseline (2014 study data)	68%	65%	80%	15%	44%
	Proposed 5-year target	27%	26%	32%	6%	18%
Emergent readers and mathematicians	Baseline (2014 study data)	12%	11%	7%	26%	19%
	Proposed 5-year target	36%	33%	21%	39%	30%
Readers and mathematicians	Baseline (2014 study data)	2%	1%	2%	4%	9%
	Proposed 5-year target	8%	4%	8%	12%	27%

Source: RTI International (2015b).

# BIBLIOGRAPHY

- Abadzi, H. (2006). *Efficient learning for the poor*. Washington, DC: The World Bank. <https://openknowledge.worldbank.org/handle/10986/7023>
- Adolf, S. M., Catts, H. W., & Lee, J. (2010). Kindergarten predictors of second versus eighth grade reading comprehension impairments. *Journal of Learning Disabilities, 43*(4), 332–345. <http://dx.doi.org/10.1177/0022219410369067>
- Abu-Rabia, S. (2000). Effects of exposure to literary Arabic on reading comprehension in a diglossic situation. *Reading and Writing: An Interdisciplinary Journal, 13*, 147–157.
- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Adolf, S. M., Catts, H. W., & Lee, J. (2010). Kindergarten predictors of second versus eighth grade reading comprehension impairments. *Journal of Learning Disabilities, 43*(4), 332–345. <http://dx.doi.org/10.1177/0022219410369067>
- Armbruster, B. B., Lehr, F., & Osborn, J. (2003). *Put reading first: The research building blocks of reading instruction*. Washington, DC: Center for the Improvement of Early Reading Achievement (CIERA).
- August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners*. Prepared by the Center for Applied Linguistics and SRI International for the Institute of Education Sciences and the Office of English Language Acquisition, U.S. Department of Education; and the National Institute of Child Health and Human Development. Washington, DC: Lawrence Erlbaum Associates and the Center for Applied Linguistics.
- Ayari, S. (1996). Diglossia and illiteracy in the Arab world. *Language, Culture and Curriculum, 9*, 243–253.
- Badian, N. A. (2001). Phonological and orthographic processing: Their roles in reading prediction. *Annals of Dyslexia, 51*, 179–202.
- Batchelder, K., Betts, K., Mulcahy-Dunn, A. & Stern, J. (2015). Lot quality assurance sampling (LQAS) pilot in Tanzania: Final report. Prepared for USAID under the EdData II project, Task Order No. AID-OAA-12-BC-00003 (RTI Task 20, Activity 5). Research Triangle Park, NC: RTI International.
- Braun, H., & Kanjee, A. (2006). Using assessment to improve education in developing nations. In H. Braun, A. Kanjee, E. Bettinger, & M. Kremer (Eds.), *Improving education through assessment, innovation, and evaluation* (pp. 1–46). Cambridge, MA: American Academy of Arts and Sciences. Retrieved from <https://www.amacad.org/publications/braun.pdf>

- Bulat, J., Brombacher, A., Slade, T., Iriondo-Perez, J., Kelly, M., & Edwards, S. (2014). *Projet d'Amélioration de la Qualité de l'Éducation (PAQUED): 2014. Endline report of Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA)*. Prepared for USAID under Contract No. AID-623-A-09-00010. Washington, DC: Education Development Center and RTI International.
- Center for Global Development. (2006). *When will we ever learn? Improving lives through impact evaluation*. [www.cgdev.org/files/7973\\_file\\_WillWeEverLearn.pdf](http://www.cgdev.org/files/7973_file_WillWeEverLearn.pdf)
- Chabbott, C. (2006). *Accelerating early grades reading in high priority EFA Countries: A desk review*. <http://www.equip123.net/docs/E1-EGRinEFACountriesDeskStudy.pdf>
- Chall, J. (1996). *Stages of reading development* (2nd ed.). Fort Worth, TX: Harcourt-Brace.
- Chiappe, P., Siegel, L., & Wade-Woolley, L. (2002). Linguistic diversity and the development of reading skills: A longitudinal study. *Scientific Studies of Reading*, 6(4), 369–400.
- Clay, M. (1993). *An observation survey of early literacy achievement*. Ortonville, MI: Cornucopia Books.
- Collins, P., & Messaoud-Galusi, S. (2012). *Student performance on the Early Grade Reading Assessment (EGRA) in Yemen* [English version; also available in Arabic]. Report prepared for USAID under the EdData II project, Task Order EHC-E-07-04-00004-00 (RTI Task 7). Research Triangle Park, NC: RTI International. [http://pdf.usaid.gov/pdf\\_docs/PNADZ047.pdf](http://pdf.usaid.gov/pdf_docs/PNADZ047.pdf)
- Coltheart M., Rastle K., Perry C., Langdon R., & Ziegler J. C. (2001). DRC: a dual-route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204–256.
- Crouch, L., & Korda, M. (2008). *EGRA Liberia: Baseline assessment of reading levels and associated factors*. Report prepared for the World Bank under Contract No. 7147768. Research Triangle Park, NC: RTI International.
- Crouch, L., & Winkler, D. (2008). Governance, management, and financing of Education for All: Basic frameworks and case studies. Background paper commissioned for the *Education for All global monitoring report 2009: Governance, management and financing of education for all*. Research Triangle Park, NC: RTI International. [unesdoc.unesco.org/images/0017/001787/178719e.pdf](http://unesdoc.unesco.org/images/0017/001787/178719e.pdf)
- Cunningham, P.M., & Allington, R. L. (2015). *Classrooms that work: They can all read and write* (6th ed.). Boston, MA: Pearson.
- Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of Learning Disabilities*, 39(6), 507–514. <http://dx.doi.org/10.1177/00222194060390060301>

- Darney, D., Reinke, W. M., Herman, K. C., Stormont, M., & Jalongo, N. S. (2013). Children with co-occurring academic and behavior problems in first grade: Distal outcomes in twelfth grade. *Journal of School Psychology, 51*(1), 117–128. <http://dx.doi.org/10.1016/j.jsp.2012.09.005>
- Denton, C. A., Ciancio, D. J., & Fletcher, J. M. (2006). Validity, reliability, and utility of the observation survey of early literacy achievement. *Reading Research Quarterly, 41*(1), 8–34.
- Denton, C. A., Hasbrouck, J. E., Weaver, L. R., & Riccio, C. A. (2000). What do we know about phonological awareness in Spanish? *Reading Psychology, 21*, 335–352.
- Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development, 40*, 315–322. <http://dx.doi.org/10.1016/j.ijedudev.2014.11.004>
- du Plessis, J., El-Ashry, F., & Tietjen, K. (Forthcoming). Oral reading assessments in Yemen: Turning bad news into a national reform. In *Understanding what works in oral reading assessments*. Montreal: UNESCO Institute for Statistics (UIS).
- Dynamic Measurement Group, Inc. (2010). *DIBELS® Next benchmark goals and composite score*. <https://dibels.org/papers/DIBELSNextBenchmarkGoals.pdf>
- Ehri, L. C. (1998). Grapheme-phoneme knowledge is essential for learning to read words in English. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 3–40). Mahwah, NJ: Lawrence Erlbaum Associates.
- Ehri, L. C., & Wilce, L. S. (1985). Movement into reading: Is the first stage of printed word learning visual or phonetic? *Reading Research Quarterly, 20*(2), 163–179.
- Ferguson, C. A. (1959). Diglossia. *Word, 15*, 325–340.
- Ferrara, S., Perie, M., & Johnson, E. (2008). Matching the judgmental task with standard setting panelist expertise: The item-descriptor (ID) matching method. *Journal of Applied Testing Technology, 9*(1), 1–22.
- Filmer, D., Hasan, A., & Pritchett, L. (2006). *A millennium learning goal: Measuring real progress in education*. Washington, DC: World Bank. Retrieved from <http://dx.doi.org/10.2139/ssrn.982968>
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.) New York: John Wiley.
- Fuchs, L., Fuchs, D., Hosp, M. K., & Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256.
- Gambrell, L. B., & Morrow, L. M. (Eds.). (2014). *Best practices in literacy instruction* (5th ed.). New York, NY: Guilford.
- Glick, P., & Sahn, D. E. (2010). Early academic performance, grade repetition, and school attainment in Senegal: A panel data analysis. *The World Bank Economic Review, 24*(1), 93–120.

- Goikoetxea, E. (2005). Levels of phonological awareness in preliterate and literate Spanish-speaking children. *Reading and Writing, 18*, 51–79.
- Good, R. H., Simmons, D. C., & Smith, S. (1998). Effective academic intervention in the United States: Evaluating and enhancing the acquisition of early reading skills. *School Psychology Review, 27*, 45–56.
- Goswami, U. (2008). The development of reading across languages. *Annals of the New York Academy of Sciences, 1145*, 1–12.
- Gove, A., & Cvelich, P. (2011). *Early reading: Igniting education for all. A report by the Early Grade Learning Community of Practice* (rev. ed). Research Triangle Park, NC: RTI International.  
<http://www.rti.org/publications/abstract.cfm?pubid=17099>
- Gove, A., & Wetterberg, A. (2011). The Early Grade Reading Assessment: An introduction. In A. Gove & A. Wetterberg (Eds.), *The Early Grade Reading Assessment: Applications and interventions to improve basic literacy* (pp. 1–37). Research Triangle Park, NC: RTI Press. <http://www.rti.org/pubs/bk-0007-1109-wetterberg.pdf>
- Gove, A., & Wetterberg, A. (Eds.). (2011). *The Early Grade Reading Assessment: Applications and interventions to improve basic literacy*. Research Triangle Park, NC: RTI Press. <http://www.rti.org/pubs/bk-0007-1109-wetterberg.pdf>
- Hamilton, L. S., Stetcher, B. M., & Yuan, K. (2008). *Standards-based reform in the United States: history, research, and future directions*. Prepared under National Science Foundation Grant No. REC-0228295. Santa Monica, CA: RAND Corporation.  
[http://www.rand.org/content/dam/rand/pubs/reprints/2009/RAND\\_RP1384.pdf](http://www.rand.org/content/dam/rand/pubs/reprints/2009/RAND_RP1384.pdf)
- Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Research Report 94-4). Iowa City, IA: ACT.
- Hanushek, E. A., & Woessman, L. (2009). *Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation*. Working Paper 14633. Cambridge, MA: National Bureau of Economic Research.
- Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*(7), 636–644.
- Hirsch Jr., E. D. (2003). Reading comprehension requires knowledge of words and the world: Scientific insights into the fourth-grade slump and the nation's stagnant comprehension scores. *American Educator* (Spring), 10–44.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*, 127–160.
- Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher, 58*(8), 702–714.

- Institute of Education Sciences, National Center for Education Statistics [US]. (n.d.). *International comparisons in fourth-grade reading literacy: Reading literacy by benchmarks* (Web page). <http://nces.ed.gov/pubs2004/pirlspub/5.asp>
- Jakobsen, R. (1960). Closing statements: Linguistics and poetics. In T. A. Sebeok (Ed.), *Style in language* (pp. 350–377). Cambridge, MA: MIT Press.
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology* 80(4), 437–447.
- Juel, C. (1991). Beginning reading. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (pp. 759–788). New York: Longman.
- Kamhi, A.G., & Catts, H. W. (1991). Language and reading: Convergences, divergences, and development. In A. G. Kamhi & H. W. Catts (Eds.), *Reading disabilities: A developmental language perspective* (pp. 1–34). Toronto, Ontario, Canada: Allyn & Bacon.
- Kanjee, A. (2009). *Assessment overview* [Presentation]. Prepared for the first READ Global Conference, "Developing a Vision for Assessment Systems," Moscow, October 1, 2009.  
[http://www.worldbank.org/content/dam/Worldbank/document/Program/READ/Events/READ-conference-2009/READ\\_GC\\_Presentation\\_5\\_AKanee\\_Eng.pdf](http://www.worldbank.org/content/dam/Worldbank/document/Program/READ/Events/READ-conference-2009/READ_GC_Presentation_5_AKanee_Eng.pdf)
- Kleinman, L., Leidy, N. K., Crawley, J., Bonomi, A., & Schoenfeld, P. (2001). A comparative trial of paper-and-pencil versus computer administration of the quality of life in reflux and dyspepsia (QOLRAD) questionnaire. *Medical Care* 39, 181–189.
- Kochetkova, E., & Dubeck, M. (In press). Assessment in schools. Chapter in *Understanding what works in oral reading assessments*. Montreal: UNESCO Institute for Statistics (UIS).
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer-Verlag.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- LaTowsky, R. (2014). *Towards possible early grade reading benchmarks for the West Bank* (Presentation slides). Prepared for USAID under the Education Data for Decision Making (EdData II) project, Measurement and Research Support to Education Strategy Goal 1, Task Order No. AID-OAA-12-BC-00003 (RTI Task 20). Research Triangle Park, NC: RTI International.  
<https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=778>

- LaTowsky, R.J., Cummiskey, C., & Collins, P. (2013). *Egypt grade 3 Early Grade Reading Assessment baseline. Draft for review and comment*. Prepared for USAID under the Education Data for Decision Making (EdData II) project, Data for Education Programming in Asia and the Middle East (DEP-AME) task order, Contract No. AID-278-BC-00019. Research Triangle Park, NC: RTI International.
- Linan-Thompson, S., & Vaughn, S. (2004). *Research-based methods of reading instruction: Grades K-3*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Linan-Thompson, S., & Vaughn, S. (2007). *Research-based methods of reading instruction for English-language learners: Grades K-4*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Lonigan, C., Wagner, R., Torgesen, J. K., & Rashotte, C. (2002). *Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP)*. Tallahassee: Department of Psychology, Florida State University.
- Management Systems International (MSI). (2014). Early Grade Reading Assessment baseline report. Balochistan province. Prepared for USAID under the Monitoring and Evaluation Program (MEP), Contract No. AID-391-C-13-00005. Washington, DC: MSI. [http://pdf.usaid.gov/pdf\\_docs/PA00KB9N.pdf](http://pdf.usaid.gov/pdf_docs/PA00KB9N.pdf)
- Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in grades K-2 in Spanish-speaking English language learners. *Learning Disabilities Research and Practice, 19*(4), 214-224.
- Marsick, V. J., & Watkins, K. E. (2001). Informal and incidental learning. *New Directions for Adult and Continuing Education, 89*, 25-34. <http://tecfa.unige.ch/staf/staf-kborer/Memoire/incidentallearning/incidentallearning.pdf>
- McBride-Chang, C. & Ho, C. S.-H. (2005). Predictors of beginning reading in Chinese and English: A 2-year longitudinal study of Chinese kindergarteners. *Scientific Studies of Reading, 9*, 117-144.
- McBride-Chang, C., & Kail, R. (2002). Cross-cultural similarities in the predictors of reading acquisition. *Child Development, 73*, 1392-1407.
- Mulcahy-Dunn, A., Valadez, J. J., Cummiskey, C., & Hartwell, A. (2013). *Report on the pilot application of lot quality assurance sampling (LQAS) in Ghana to assess literacy and teaching in primary grade 3*. Prepared for USAID under the EdData II project, Task Order No. EHC-E-07-04-00004-00 (RTI Task 7). Research Triangle Park, NC: RTI International.
- Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundation of early reading development: Evidence from a longitudinal study. *Developmental Psychology, 40*, 665-681.
- Nag, S. (2007). Early reading in Kannada: The pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading, 30*(1), 7-22.

- Nag, S. (2014). Akshara-phonology mappings: the common yet uncommon case of the consonant cluster. *Writing Systems Research*, 6, 105–119.
- Nag, S., & Perfetti, C. A. (2014). Reading and writing: Insights from the alphasyllabaries of South and Southeast Asia. *Writing Systems Research*, 6(1), 1–9.
- Nagy, W. E., & Scott, J. (2000). Vocabulary processes. In M. E. A. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr, (Eds.), *Handbook of reading research* (Vol. III, pp. 269-284). Mahwah, NJ: Erlbaum.
- Nation, K. (2005). Connections between language and reading in children with poor reading comprehension. In H. W. Catts & A. G. Kamhi (Eds.), *The connections between language and reading disabilities* (pp. 41–54). Mahwah, NJ: Erlbaum.
- National Center for Family Literacy (NCFL) [US]. (2008). *Developing early literacy: Report of the national early literacy panel. A scientific synthesis of early literacy development and implications for intervention*. Prepared under inter-agency agreement IAD-01-1701 and IAD-02-1790 between the Department of Health and Human Services and the National Institute for Literacy. Washington, DC: National Institute for Literacy.  
[https://www.nichd.nih.gov/publications/Pages/pubs\\_details.aspx?pubs\\_id=5750](https://www.nichd.nih.gov/publications/Pages/pubs_details.aspx?pubs_id=5750)
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, US Department of Health, Education and Welfare (DHEW). (1978). *Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. DHEW Pub. No. (OS) 78-0012. Washington, DC: United States Government Printing Office. [http://videocast.nih.gov/pdf/ohrp\\_belmont\\_report.pdf](http://videocast.nih.gov/pdf/ohrp_belmont_report.pdf)
- National Institute of Child Health and Human Development (NICHD). (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: NICHD.  
<https://www.nichd.nih.gov/publications/pubs/nrp/Pages/smallbook.aspx>
- Nielsen, D. (2014). *Early grade reading and math assessments in 10 countries: Dissemination and utilization of results—a review*. Report prepared for USAID under the Education Data for Decision Making (EdData II) project, Measurement and Research Support to Education Strategy Goal 1, Task Order No. AID-OAA-BC-12-00003 (RTI Task 20). Research Triangle Park, NC: RTI International.  
[http://pdf.usaid.gov/pdf\\_docs/PA00K8RP.pdf](http://pdf.usaid.gov/pdf_docs/PA00K8RP.pdf)
- Office of the United Nations Secretary-General. (2012). *Global Education First Initiative: An initiative of the United Nations Secretary-General*. New York: United Nations. [http://www.globaleducationfirst.org/files/GEFI\\_Brochure\\_ENG.pdf](http://www.globaleducationfirst.org/files/GEFI_Brochure_ENG.pdf)

- Optimal Solutions Group, LLC. (2015). *Secondary Analysis for Results Tracking (SART) data sharing manual, USAID Ed Strategy 2011–2015, Goal 1*. Prepared for USAID under the Secondary Analysis for Results Tracking (SART) project, Contract AID-OAA-C-12-00069. Location: Optimal Solutions. <https://sartdatacollection.org/images/SARTDataSharingManualFeb2015.pdf>
- Orr, D. B., & Graham, W. R. (1968). Development of a listening comprehension test to identify educational potential among disadvantaged junior high school students. *American Educational Research Journal*, 5(2), 167–180.
- Paris, S. G., & Paris, A. H. (2006). Chapter 2: Assessments of early reading. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development, 6th Edition* (Vol. 4: Child Psychology in Practice). Hoboken, New Jersey: John Wiley and Sons.
- Patrinos, H. A., & Velez, E. (2009). Costs and benefits of bilingual education in Guatemala: A partial analysis. *International Journal of Educational Development*, 29(6), 594–598.
- Perfetti, C. A., & Dunlap, S. (2008). Learning to read: General principles and writing system variations. In K. Koda & A. Zehler (Eds.), *Learning to read across languages* (pp. 13–38). Mahwah, NJ: Erlbaum.
- Piper, B., & Korda, M. (2010). *EGRA Plus: Liberia. Program evaluation report*. Prepared for USAID/Liberia under the Education Data for Decision Making (EdData II) project, Early Grade Reading Assessment (EGRA): Plus Project, Task Order No. EHC-E-06-04-00004-00 (RTI Task 6). Research Triangle Park, NC: RTI International. [http://pdf.usaid.gov/pdf\\_docs/pdacr618.pdf](http://pdf.usaid.gov/pdf_docs/pdacr618.pdf)
- Piper, B., & Mugenda, A. (2014). *USAID/Kenya Primary Math and Reading (PRIMR) Initiative: Endline impact evaluation*. Prepared under the USAID EdData II project, Task Order No. AID-623-M-11-00001 (RTI Task 13). Research Triangle Park, NC: RTI International. [http://pdf.usaid.gov/pdf\\_docs/pa00k27s.pdf](http://pdf.usaid.gov/pdf_docs/pa00k27s.pdf)
- Piper, B., & Zuilkowski, S. S. (2015). The role of timing in assessing oral reading fluency and comprehension in Kenya. *Language Testing* [online publication]. <http://dx.doi.org/10.1177/0265532215579529>
- Prodigy Systems. (2011). *EGRA Yemen with iProSurveyor* [Presentation slides]. Sana'a: Prodigy Systems.
- Pouzevara, S. Costello, M. Banda, O. (2012). *Malawi National Early Grade Reading Assessment survey. Final assessment – November 2012*. Prepared for USAID under the Malawi Teacher Professional Development Support (MTPDS) program, Contract No. EDH-I-00-05-00026-02; Task Order No. EDH-I-04-05-00026-00. Washington, DC: Creative Associates International, RTI International, and Seward, Inc. [http://pdf.usaid.gov/pdf\\_docs/PA00JB9R.pdf](http://pdf.usaid.gov/pdf_docs/PA00JB9R.pdf)
- Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M.S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, 2, 31–74.

- Roth, F. P., Speece, D. L., & Cooper, D. H. (2002). A longitudinal analysis of the connection between oral language and early reading. *Journal of Educational Research, 95*, 259–272.
- RTI International. (2008). *Early grade reading Kenya: Baseline assessment. Analyses and implications for teaching interventions design. Final report*. Prepared for USAID under the EdData II project, Task Order No. EHC-E-01-04-00004-00 (RTI Task 4). Research Triangle Park, NC: RTI International. [http://pdf.usaid.gov/pdf\\_docs/PNADL212.pdf](http://pdf.usaid.gov/pdf_docs/PNADL212.pdf)
- RTI International. (2011). *EGRA Plus: Liberia. Final report: October 2008–January 2011*. Prepared for USAID/Liberia under the EdData II Project, Task Order No. EHC--E-06-04-00004-00 (RTI Task 6). Research Triangle Park, NC: RTI International. [http://pdf.usaid.gov/pdf\\_docs/PNADZ817.pdf](http://pdf.usaid.gov/pdf_docs/PNADZ817.pdf)
- RTI International. (2014a). *Codebook for EGRA and EGMA* [Excel spreadsheet]. Research Triangle Park, NC: RTI. Retrieved from <https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=389>
- RTI International. (2014b). *USAID/Kenya Primary Math and Reading (PRIMR) Initiative: Final report*. Prepared for USAID under the EdData II project, Task Order No. AID-623-M-11-00001. Research Triangle Park, NC: RTI. [http://pdf.usaid.gov/pdf\\_docs/PA00K282.pdf](http://pdf.usaid.gov/pdf_docs/PA00K282.pdf)
- RTI International. (2015a). EGRA tracker. Prepared for USAID under the EdData II project, Contract No. EHC-E-00-04-00004-00. Research Triangle Park, NC: RTI. <https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=188>
- RTI International. (2015b). *Proposing benchmarks and targets for early grade reading and mathematics in Zambia* [3-page brief]. Prepared for MOGE, USAID, and DFID under the USAID Education Data for Decision Making (EdData II) project, Measurement and Research Support to Education Strategy Goal 1, Task Order No. AID-OAA-12-BC-00003 (RTI Task 20). Research Triangle Park, NC: RTI International.
- RTI International & International Rescue Committee (IRC). (2011). *Guidance notes for planning and implementing EGRA*. Research Triangle Park, NC: RTI and IRC. <https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=318>
- Saiegh-Haddad, E. (2003). Linguistic distance and initial reading acquisition: the case of Arabic diglossia. *Applied Psycholinguistics, 24*, 115–135.
- Scanlon, D. M., Gelzheiser, L. M., Vellutino, F. R., Schatschneider, C., & Sweeney, J. M. (2008). Reducing the incidence of early reading difficulties: Professional development for classroom teachers versus direct interventions for children. *Learning and Individual Differences, 18*(3), 346–359. <http://dx.doi.org/10.1016/j.lindif.2008.05.002>
- Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143–174.

- Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, 134(4), 584–615.
- Share, D. L., Jorm, A., Maclearn, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Education Psychology*, 76, 1309–1324.
- Share, D. L., & Leikin, M. (2004). Language impairment at school entry and later reading disability: Connections at lexical versus supralephical levels of reading. *Scientific Studies of Reading*, 8, 87–110.
- Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement*, 42, 309–330.
- Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Prepared on behalf of the Committee on the Prevention of Reading Difficulties in Young Children under Grant No. H023S50001 of the National Academy of Sciences and the U.S. Department of Education. Washington, DC: National Academy Press.
- Snow, C., & the RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Research prepared for the Office of Educational Research and Improvement (OERI), U.S. Department of Education. Santa Monica, CA: RAND Corporation.
- Spencer, L. H., & Hanley, J. R. (2003). Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales. *British Journal of Psychology*, 94(1), 1–28.
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21, 360–406.
- Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: Guilford Press.
- Strigel, C. (2012). *Tangerine™—Electronic data collection tool for early reading and math assessments. January 2012 – Kenya field trial report: Summary*. Research Triangle Park, NC: RTI International. [www.rti.org/files/tangerine\\_report\\_0112.pdf](http://www.rti.org/files/tangerine_report_0112.pdf)
- Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology*, 40(1), 7–26. [http://dx.doi.org/10.1016/s0022-4405\(01\)00092-9](http://dx.doi.org/10.1016/s0022-4405(01)00092-9)
- United Nations. (2015). *The Millennium Development Goals report 2015*. New York: United Nations. [http://www.un.org/millenniumgoals/2015\\_MDG\\_Report/pdf/MDG%202015%20rev%20\(July%201\).pdf](http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf)
- United Nations Development Programme (UNDP). (2015). *Sustainable Development Goals (SDGs)* [Web page]. Retrieved from <http://www.undp.org/content/undp/en/home/mdgoverview/post-2015-development-agenda.html>

- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2014). *Education for All Global Monitoring Report 2013/4. Teaching and learning: Achieving quality for all*. Paris: UNESCO. <http://en.unesco.org/gem-report/report/2014/teaching-and-learning-achieving-quality-all#sthash.n1q0vitl.dpbs>
- United States Agency for International Development (USAID). (2012). *How-to note: Preparing evaluation reports*. Monitoring and Evaluation Series, No. 1, Version 1.0. Washington, DC: USAID. [https://www.usaid.gov/sites/default/files/documents/1870/How-to-Note\\_Preparing-Evaluation-Reports.pdf](https://www.usaid.gov/sites/default/files/documents/1870/How-to-Note_Preparing-Evaluation-Reports.pdf)
- Valadez, J. J., Mulcahy-Dunn, A., & Sam-Bossman, E. (2014). *Using lot quality assurance sampling to monitor impact of early grade reading programs* [87-slide training presentation plus handouts]. Prepared under the EdData II project, Task Order No. AID-OAA-12-BC-00003 (RTI Task 20), for a USAID-hosted webinar based in Washington, DC, July 9–10, 2014. Research Triangle Park, NC: RTI International. <https://www.eddataglobal.org/reading/index.cfm?fuseaction=pubDetail&ID=602>
- Vaughn, S., & Linan-Thompson, S. (2004). *Research-based methods of reading instruction grades K-3*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Wagner, D.A. (2011). *Smaller, quicker, cheaper: Improving learning assessments for developing countries*. Paris: UNESCO International Institute of Educational Planning (IIEP) and Fast Track Initiative/World Bank. <http://unesdoc.unesco.org/images/0021/002136/213663e.pdf>
- Wagner R. K., Torgesen J. K., & Rashotte C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bi-directional causality from a latent variable longitudinal study. *Developmental Psychology*, 30, 73–87.
- Walther, B., Hossin, S., Townend, J., Abernethy, N., Parker, D., & Jeffries, D. (2011). Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. *PLoS One*, 6(9), e25348. <http://dx.doi.org/10.1371/journal.pone.0025348>
- Wang, M., Park, Y., & Lee, K. R. (2006). Korean-English biliteracy acquisition: Cross-language phonological and orthographic transfer. *Journal of Education Psychology*, 98, 148–158.
- What Works Clearinghouse. (2015). *Procedures and standards handbook, version 3.0*. Washington, DC: Institute of Education Sciences, US Department of Education. [http://ies.ed.gov/ncee/wwc/pdf/reference\\_resources/wwc\\_procedures\\_v3\\_0\\_standards\\_handbook.pdf](http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf)
- World Bank. (2015a). *EdStats dashboards: Learning outcomes dashboard* [Web page]. Washington, DC: World Bank. [http://datatopics.worldbank.org/education/wDashboard/tbl\\_index.aspx](http://datatopics.worldbank.org/education/wDashboard/tbl_index.aspx)

- World Bank. (2015b). *Learning outcomes* [Web page]. Washington, DC: World Bank. <http://go.worldbank.org/GOBJ17VV90>
- World Bank: Independent Evaluation Group. (2006). *From schooling access to learning outcomes—An unfinished agenda: An evaluation of World Bank support to primary education*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/7083>
- Yesil-Dağlı, Ü. (2011). Predicting ELL students' beginning first grade English oral reading fluency from initial kindergarten vocabulary, letter naming, and phonological awareness skills. *Early Childhood Research Quarterly*, 26(1), 15–29.
- Yovanoff, P., Duesbery, L., Alonzo, J., & Tindall, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement, Fall*, 4–12.
- Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ: Educational Testing Service. [https://www.ets.org/Media/Research/pdf/Cut\\_Scores\\_Primer.pdf](https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf)
- Zimmerman, R. (2008). *Digital data collection demonstration white paper. A comparison of two methodologies: Digital and paper-based*. Prepared for USAID under the Educational Quality Improvement Program 1 (EQUIP1), Cooperative Agreement No. GDG-A-00-03-00006-00. Washington, DC: American Institutes for Research. <http://www.equip123.net/docs/e1-DigitalDataCollection.pdf>
- Zorzi M. (2010). The connectionist dual process (CDP) approach to modelling reading aloud. *European Journal of Cognitive Psychology*, 22, 836–860.

# ANNEX A: RECOMMENDATIONS AND CONSIDERATIONS FOR CROSS-LANGUAGE COMPARISONS

## A.1 Recommendations for the Nature of Writing Systems

To help make reasonable cross-linguistic comparisons, those adapting the EGRA tool must possess in-depth understanding of characteristics of the writing systems of the languages in question.

To improve the quality of cross-linguistic comparisons, one must know if the writing system of the language in question is morphosyllabic, syllabic, alphasyllabic, or alphabetic (Latin or non-Latin alphabetic).

The following guidelines are recommended in accordance with the type of language.

### A.1.1 Roman-Alphabetic Languages

Within Roman-alphabetic languages:

1. Know if the orthographic depth of the language in question is shallow (transparent) or deep (opaque).
  - Research suggests that children who learn to read in shallow orthographies may learn to decode more quickly than those who learn to read in deep orthographies (Spencer & Hanley, 2003). Depth of the orthography is also related to how quickly and easily comprehension is attained (e.g. Share, 2008).
2. Know the syllable structure of the language in question.
  - Languages with complex syllables (e.g., consonant-vowel combinations such as *ccvcc*, as in “starts”) take longer to learn to read than languages in which simple syllables (e.g., *cv*, as in “mesa”) predominate.
3. Know that word length influences cross-linguistic comparisons.
  - Shorter words are recognized more quickly than longer words. For example, compare agglutinative languages, which connect several morphemes, with non-agglutinative languages.
4. Know that the written markings for tonal languages can influence comprehension, while this is unimportant for non-tonal languages.

## **A.2 Recommendations for Oral Language**

Regardless of the desire to make cross-linguistic comparisons, all adaptations of EGRA must consider multiple aspects of oral language, such as: differences in dialects or the presence of diglossia, the clarity of directions, levels of difficulty of the contents of the phonological awareness, listening, and vocabulary subtasks.

For those focusing on cross-linguistic comparisons, it is particularly important to:

1. Ensure that oral reading passages in different languages have a similar level of difficulty.
2. Ensure that vocabulary words are measuring the same word meaning or construct in both languages.

## **A.3 Recommendations for Print and Orthographic Knowledge**

The content for subtasks designed to measure print and orthographic knowledge can be controlled so that there is some comparability across languages.

Cross-linguistic comparisons would track the rate and accuracy with which students being tested in different languages recognized items appropriate for that grade level, as determined by their frequency in existing grade-level texts.

## **A.4 Recommendations for Reading Connected Text**

Ensuring technical adequacy and basic comparability of connected-text reading passages in multiple-language administrations requires several considerations:

1. The passage is original writing prepared specifically for the assessment.
2. The passage addresses an age-appropriate topic in a familiar text structure, to minimize the influence of background knowledge on comprehension.
3. To best compare across languages, texts in both languages contain common story elements and topics familiar in both language groups.
4. The passage avoids the use of ambiguous words, such as:
  - A word that, spelled in one way, can represent more than one meaning (e.g., “wind” in English).
  - A word that can use more than one spelling to represent one meaning.

## **A.5 Recommendations for Second Language/Multilingual Learners**

1. When comparisons are made between languages, ensure that they are made between the same “language classification.” For example, if a test is conducted among a group of English monolinguals or English first-language speakers, then comparisons are not made to English second-language (or later language) groups.
2. Simultaneous language acquisition (or learning two or more languages from birth or an early age) is possible, so a child may have two first languages.

3. There is potential for “transfer” of skills (that is, most decoding skills can be transferred among similar writing systems) when children are reading in an additional or nonnative language.
4. If a child is learning in a second (or later) language without adequate instruction in the first language, interpretation of results reflects this. It is likely to take children much longer to reach reading proficiency in these cases.

## References for Annex A

- Abu-Rabia, S. (2000). Effects of exposure to literary Arabic on reading comprehension in a diglossic situation. *Reading and Writing: An Interdisciplinary Journal*, 13, 147–157.
- Ayari, S. (1996). Diglossia and illiteracy in the Arab world. *Language, Culture and Curriculum*, 9, 243–253.
- Ferguson, C. A. (1959). Diglossia. *Word*, 15, 325–340.
- Nag, S. (2007). Early reading in Kannada: The pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading*, 30(1), 7–22.
- Nag, S. (2014). Akshara-phonology mappings: the common yet uncommon case of the consonant cluster. *Writing Systems Research*, 6, 105–119.
- Nag, S., & Perfetti, C. A. (2014). Reading and writing: Insights from the alphasyllabaries of South and Southeast Asia. *Writing Systems Research*, 6(1), 1–9.
- Saiegh-Haddad, E. (2003). Linguistic distance and initial reading acquisition: the case of Arabic diglossia. *Applied Psycholinguistics*, 24, 115–135.
- Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, 134(4), 584–615.
- Spencer, L. H., & Hanley, J. R. (2003). Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales. *British Journal of Psychology*, 94(1), 1–28.

# ANNEX B: SAMPLE ASSESSOR TRAINING AGENDA

## Training EGRA Data Collectors

Day & Time	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
<b>Daily Objectives:</b>	<ul style="list-style-type: none"> <li>Understand purpose of EGRA</li> <li>Be able to apply administration and scoring rules on paper</li> </ul>	<ul style="list-style-type: none"> <li>Understand tablet functions and administration</li> <li>Be able to upload data</li> </ul>	<ul style="list-style-type: none"> <li>Improve test administration skills</li> <li>Become familiar with questionnaire administration</li> </ul>	<ul style="list-style-type: none"> <li>Polish EGRA administration skills and scoring accuracy</li> </ul>	<ul style="list-style-type: none"> <li>Polish EGRA administration skills and scoring accuracy</li> </ul>	<ul style="list-style-type: none"> <li>Supervisor training</li> <li>Team preparations</li> </ul>
8:30-9:00 a.m.	<ul style="list-style-type: none"> <li>Welcome/introductions</li> </ul>	<ul style="list-style-type: none"> <li>Review of Day 1</li> </ul>	School visit 1: EGRA practice	School visit 2: EGRA + questionnaires	School visit 3: EGRA + questionnaires	<ul style="list-style-type: none"> <li>Supervisor training</li> <li>Team preparations for data collection</li> </ul>
9:00-10:30 a.m.	<ul style="list-style-type: none"> <li>Overview of EGRA: purpose, instrument content</li> <li>Purpose of EGRA in this context</li> </ul>	<ul style="list-style-type: none"> <li>Overview of basic tablet functions</li> </ul>				
10:30-11:00 a.m.	<i>Break</i>	<i>Break</i>				
11:00-1:00 p.m.	<ul style="list-style-type: none"> <li>Instrument overview</li> <li>Demonstration and practice of subtasks</li> </ul>	<ul style="list-style-type: none"> <li>Practice EGRA on tablets (small groups)</li> </ul>				
1:00-2:00 p.m.	<i>Lunch</i>					
2:00-3:30 p.m.	<ul style="list-style-type: none"> <li>Continued demonstration and practice of subtasks</li> <li>Pupil questionnaire</li> </ul>	<ul style="list-style-type: none"> <li>Tablet functionality issues</li> <li>Uploading data</li> </ul>	<ul style="list-style-type: none"> <li>School visit debrief</li> <li><i>Additional survey instruments if administered</i></li> </ul>	<ul style="list-style-type: none"> <li>School visit debrief</li> <li>Discuss IRR 2 results</li> <li>Practice EGRA on tablets in pairs (key tasks/issues)</li> </ul>	<ul style="list-style-type: none"> <li>School visit debrief</li> <li>Discuss IRR 2 results</li> <li><b>Data collection logistics</b></li> </ul>	

Day & Time	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
3:30-3:45 p.m.	<i>Break</i>					
3:45-5:30 p.m.	<ul style="list-style-type: none"> <li>Continued whole and small-group practice and correction</li> </ul>	<ul style="list-style-type: none"> <li>EGRA sampling procedures</li> <li>School visit logistics</li> </ul>	<ul style="list-style-type: none"> <li>Practice EGRA on tablets in pairs (key tasks/issues)</li> <li><b>Performance Assessment (IRR) 1</b></li> <li>Review school visit logistics</li> </ul>	<ul style="list-style-type: none"> <li><b>Performance Assessment (IRR) 2</b> <i>Additional survey instruments if administered</i></li> </ul>	<ul style="list-style-type: none"> <li><b>Performance Assessment (IRR) 3</b></li> </ul>	

The number of training days and content of sessions greatly depends on the number of instruments that will be administered (EGRA plus other questionnaires, or in multiple languages), the number of assessors to train, and their level of experience. If assessors will learn to administer EGRA in two languages, more time will need to be spent training them on EGRA. As a result, it is recommended that the number of school visits be reduced to two, to provide more time during the workshop for them to learn the instrument.

# ANNEX C: DATA ANALYSIS AND STATISTICAL GUIDANCE FOR MEASURING ASSESSORS' ACCURACY

This annex provides details about managing the data collected for gauging assessors' accuracy, including some related statistical terminology and guidance.

## C.1 Data Preparation

**Exhibit C-1** is an example that shows (indicated by the shaded cells) at an item level where the Gold Standard and mode differed. If this occurs, the training team investigates why. Possible explanations could be that the Gold Standard was inaccurate, there was a problem with the instrument, or there was an issue with the trainees' interpretation of this item and it is the focus of further training.

**Exhibit C-1: Example of Microsoft Excel output comparing Gold Standard with the modal assessor response**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	enumerator	non_word_time_remain	non_word_attempted	non_word1	non_word2	non_word3	non_word4	non_word5	non_word6	non_word7	non_word8	non_word9	non_word10	non_word11	non_word12	non_word13	non_word14
2	GoldStdirr1	0	41	0	1	1	1	1	0	1	1	1	0	1	0	1	1
3	mode	0	41	1	1	0	1	1	1	1	1	1	0	1	1	1	1
4	mode vs. GS	.	.	!	.	!	.	.	!	.	.	.	.	.	!	.	.
5																	
6	aloreirr1	0	41	0	1	1	1	1	0	1	1	1	0	1	0	1	1
7	apanjirr1	0	42	0	1	1	1	1	0	1	1	1	0	1	0	1	1
8	ashooirr1	0	42	0	1	1	1	1	0	1	1	1	0	1	0	1	1
9	dmtitirr1	0	40	0	1	1	1	1	0	1	1	1	0	1	0	1	1
10	hseleirr1	0	41	0	1	1	1	1	0	1	1	1	0	1	0	1	1
11	ikiwairr1	0	41	0	1	1	1	1	0	1	1	1	0	1	0	1	1
12	jmasairr1	0	41	0	1	1	1	1	0	1	1	1	0	1	0	1	1
13	jurasirr1	0	41	1	1	0	1	1	1	1	1	1	0	1	1	1	1
14	kkahairr1	0	42	0	1	0	1	1	1	1	1	1	0	1	1	1	1

Verify the Gold Standard responses by comparing with modal response of assessors.

## C.2 Data Analysis

Percent agreement by assessor is then calculated by subtask. This measure is the agreement between the assessor's evaluation of the child and the correct evaluation of the child. To calculate each assessor's score (for each subtask and for the assessment as a whole), the training leader tallies the number of agreements with the Gold Standard and express this a percentage of the number of items in the subtask/assessment, as shown in **Exhibit C-2**.

**Exhibit C-2: Example of Microsoft Excel output calculating percent agreement with Gold Standard, by subtask**

enumerator	Non word	non_word_time_remain	non_word_attempted	non_word1	non_word2	non_word3	non_word4	non_word5	non_word6	non_word7	non_word8	non_word9	non_word10	non_word11	non_word12	non_word13	non_word14
Average	88%	95%	59%	32%	100%	14%	100%	100%	0%	100%	100%	100%	73%	100%	0%	100%	95%
aloreirr1	91%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
apanjirr1	81%	1	0	0	1	0	1	1	0	1	1	1	1	1	0	1	1
ashooirr1	75%	1	0	0	1	0	1	1	0	1	1	1	1	1	0	1	0
dmtitirr1	89%	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1	1
hseleirr1	91%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
ikiwairr1	91%	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1
jmasairr1	89%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
jurairr1	91%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
kkahairr1	89%	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1	1
lkayoirr1	85%	1	0	1	1	0	1	1	0	1	1	1	0	1	0	1	1
mkyejirr1	79%	1	0	0	1	1	1	1	0	1	1	1	0	1	0	1	1
mdolirr1	93%	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1
mpaziirr1	91%	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1
mramairr1	91%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
nkibonairr1	79%	0	0	1	1	0	1	1	0	1	1	1	0	1	0	1	1

Using a formula, the calculation is made as follows:

$$\text{Assessor subtask score(\%)} = \frac{\text{number of agreements with the Gold Standard}}{\text{number of items in the subtask}}$$

The item-level average agreement can also be calculated across the assessors using the formula:

$$\text{Item level agreement (\%)} = \frac{\text{\# of agreements with the Gold Standard for the item}}{\text{number of responses (assessors) for the item}}$$

If the Gold Standard has missing items because the “child” did not complete all the items for a subtask, the agreement results by assessor also include agreement with the missing items.

For timed subtasks such as oral reading fluency and correct letter sounds per minute, if a child completes the subtask within the allotted time, it is important for the assessor to take an accurate reading of the time the child took to complete that task. If the assessor is within 2 seconds of the Gold Standard time remaining, the assessor

is considered in agreement with the Gold Standard. Then an overall average percent agreement is calculated across all the time-remaining variables.

An overall percent agreement by assessor is an average of the subtask and time-remaining percent agreements. An overall assessment percent agreement is calculated as an average of the assessor overall percent.

Thus, the summary output is reported for each assessment and include the following:

- By assessor: Percent agreement by subtask and overall
- Overall percent agreement average
- Overall percent agreement by subtask.

### **C.3 Statistical Glossary and Definitions**

#### Raw % agreement

Measures the extent to which raters make exactly the same judgment

#### Kappa

Measures the extent to which two different ratings of the same subject could have happened by chance. Kappa values range from -1.0 to 1.0. Higher values indicate lower probability of chance agreement.

#### Intraclass correlation coefficient (ICC)

Describes the consistency of scores given to students by different raters. ICC values range from 0.0 to 1.0. Higher values indicate greater agreement among assessors.

### **C.4 Benchmarks for Assessor Agreement**

#### Raw % agreement

Due to the lack of detail that is generated solely by this statistic, no benchmark is possible. Efforts are made for assessors to have % agreement be as high as possible (as close to 100%) when assessing students. However, regardless of the % agreement, evaluators must reference the Kappa statistics to understand the quality of the % agreement statistic.

## Kappa

### OPTION 1

Source: Landis & Koch (1977)

Kappa Statistic	Strength of Agreement
less than 0.0	Poor
0.0 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

### OPTION 2

Source: Fleiss (1981)

Kappa Statistic	Strength of Agreement
Less than 0.40	Poor
0.40 to 0.75	Intermediate to Good
Greater than 0.75	Excellent

## Intraclass correlation coefficient

Source: Fleiss (1981)

Kappa Statistic	Strength of Agreement
Less than 0.40	Poor
0.40 to 0.75	Intermediate to Good
Greater than 0.75	Excellent

## References for Annex C

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.) New York: John Wiley.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.

# ANNEX D: SAMPLE CODEBOOK

Section: Demographic	Format	Label name	Label values	Variable label
Country	String	—	(Largest Geographical Variable)	In which country was the assessment given?
Project	String	—	—	Which project within the country?
Year	Integer (2000-2020)	—	—	In what year was the assessment conducted?
Month	Ordinal (1-12)	month	1 January 2 February . . . 12 December	In what month was the assessment conducted?
Date	Date format	—	—	On what date was the assessment conducted?
State	Nominal	state	country specific list (Second largest geographical variable, below Country)	In which state is the student's school located?
Region	Nominal	region	country specific list (Third largest geographical variable, below State)	In which region is the student's school located?
District	Nominal	district	country specific list (Smallest geographical variable, below Region)	In which district is the student's school located?
School_name	String	school	country specific list	What is the name of the student's school?
School_code	Integer	—	country specific list	School's code within country
EMIS	Integer	—	—	Education Management Information System code
School_type	Nominal	school_type	Set value labels according to project	What type of school does the student attend?
Treatment	Dichotomous	treatment	0 "Control" 1 "Partial Treatment" 2 "Full Treatment", replace	What level of treatment is the school receiving?
Treat_year	Ordinal (0-12)	—	—	How many years has the school been receiving the treatment?

Section: Demographic	Format	Label name	Label values	Variable label
Treat_phase	Ordinal (1-6)	treat_phase	Set value labels according to project	In which phase of the study is this treatment-school student?
Urban	Dichotomous	urban	0 Rural 1 Urban	Is the school in an urban area?
Shift	Ordinal (0-2)	shift	0 "No Shift" (Full Day) 1 Morning 2 Afternoon 3 Alternating	Does the student attend in school in shifts?
Dbl_shift	Dichotomous	yes/no	0 No 1 Yes	Does the school operate on double shifts?
Admin	Nominal	admin	country specific list	Who administered the test? (code number)
Admin_name	String	—	—	Who administered the test?
ID	String	—	Must be unique!!!!	Unique student identification number
Grade	Integer (1-8)	grade	1 first, 2 second, 3 third, 4 fourth, 5 fifth, 6 sixth, 7 seventh, 8 eighth	What is the student's grade level?
Level	Integer	—	Same as grade, but for students who are not of traditional age	For non-traditionally aged students, at what "grade" level are they learning?
Section	Integer	—	country specific list	In which grade section is the student?
Female	Dichotomous	female	0 Male 1 Female	Is the student female?
Multigrade	Dichotomous	yes/no	0 No 1 Yes	Is the student's class a multiple-grade classroom?
Teacher	Integer	teacher	Country-specific list	What is the name of the student's teacher?
Age	Integer (5-18)	—	—	How old is the student?
Start_time	Time (hh:mm)	—	—	Assessment start time?
End_time	Time (hh:mm)	—	—	Assessment end time?
Assess_time	Time (m)	—	—	Minutes taken to complete the assessment?
Language	Integer	language	use ISO 639-3 codes	Language of assessment
Consent	Dichotomous	yes/no	0 No 1 Yes	Did the participant give consent/assent to complete the assessment?