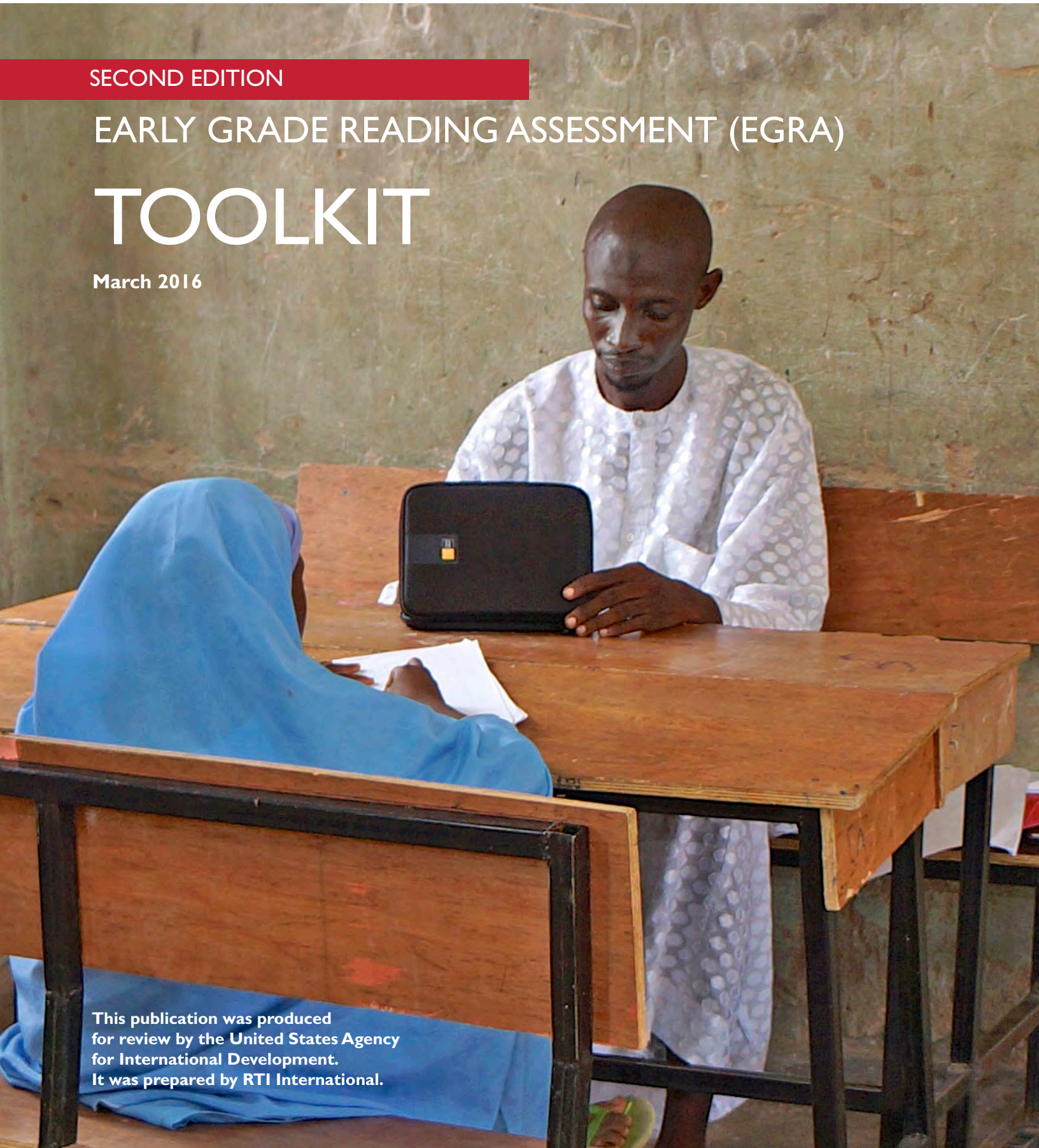
USAID
FROM THE AMERICAN PEOPLE

SECOND EDITION

EARLY GRADE READING ASSESSMENT (EGRA)

TOOLKIT

March 2016



This publication was produced for review by the United States Agency for International Development. It was prepared by RTI International.

EARLY GRADE READING ASSESSMENT TOOLKIT, SECOND EDITION

Cover photo: RTI project staff, USAID/Nigeria Northern Education Initiative

This publication was produced for the United States Agency for International Development by RTI International, 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).

RTI International. 2016. *Early Grade Reading Assessment (EGRA) Toolkit, Second Edition.* Washington, DC: United States Agency for International Development.

Copyright © 2016 by RTI International

RTI International is a registered trademark and a trade name of Research Triangle Institute.

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/. Under the Creative Commons Attribution license, you are free to copy, distribute, transmit, and adapt this work under the following conditions:

Attribution—If you copy and distribute this work in its entirety, without making changes to content or illustrations, please cite the work as follows: *Reproduced on the basis of an original work published by RTI International and licensed under the Creative Commons Attribution 4.0 International License.*

Translations—If you create a translation of this work, please use the following label on your work: *Translated from an original work published by RTI International and licensed under the Creative Commons Attribution 4.0 International License.*

Adaptations—If you create an adaptation of this work, please use the following label on your work: *This is an adaptation of an original work published by RTI International and licensed under the Creative Commons Attribution 4.0 International License.*

ACKNOWLEDGMENTS

This Early Grade Reading Assessment (EGRA) toolkit is the product of ongoing collaboration among a large community of scholars, practitioners, government officials, and education development professionals to advance the cause of early reading assessment and acquisition among primary school children in low-income countries.

Although it is not possible to recognize all of those who contributed to revising and updating the toolkit, it would not have been possible without Melissa Chiappetta of Abt Associates; Ray Adams and Juliette Mendelovits with the Australian Council for Educational Research (ACER); Pooja Reddy Nakamura and Zarko Vukmirovic at American Institutes for Research (AIR); Fathi El Ashry of Creative Associates International; Elena Vinogradova of Education Development Center, Inc. (EDC); Matt Sloan of Mathematica Policy Research; Thomaz Alvares and Abdullah Ferdous at Management Systems International (MSI); Roger Stanton at Optimal Solutions Group; Kellie Betts, Chris Cummiskey, Margaret Dubeck, Karon Harden, Simon King, Jessica Mejia, Erin Newton, Alison Pflepson, Lilly Piper, Sarah Pouezevara, and Jonathan Stern of RTI International; Elliot Friedlander and Carol da Silva at Save the Children; and Aglaia Zafeirakou at the World Bank. Also, special recognition and a thank you are owed to Truphena Choti with University Research Company (URC, LLC); Jane Benbow, formerly with URC, LLC; and other URC, LLC and Global Reading Network staff who contributed to the organization and hosting of various Early Grade Reading Assessment workshops and seminars during 2015.

Development of the original EGRA instrument would not have been possible without the support of the nongovernmental organizations and Ministry of Education EGRA Evaluation Teams of Afghanistan, Bangladesh, Egypt, The Gambia, Guyana, Haiti, Honduras, Jamaica, Kenya, Liberia, Mali, Nicaragua, Niger, Peru, Senegal, and South Africa. Our deepest gratitude goes to the teachers, the students, and their families for their participation and continued faith in the benefits of education. In repayment we will diligently seek to improve reading outcomes for all children around the world.

Amber Gove was responsible for primary authorship of the original toolkit, with contributions from Luis Crouch, Amy Mulcahy-Dunn, and Marguerite Clarke. The

second edition benefited from contributions from many additional seminar participants, leaders, and reviewers.[1]

The opinions expressed in this document are those of the authors and do not necessarily reflect the views of the United States Agency for International Development. Please direct questions or comments to Penelope Bender at pbender@usaid.gov.

[1] Names of individuals and organizations contributing directly to the second edition of the toolkit are recognized in **Annex A.**

CONTENTS

PAGE

Acknowledgments iii

List of Exhibits x

Abbreviations xii

Glossary of Terms xv

Reading-Related Terminology xv
Statistical Terms xvi
Methodological Terms xix

1 Introduction 1

1.1 Why Do We Need Early Grade Reading Assessments? 1
1.1.1 Why Assess **Reading**? 2
1.1.2 Why Assess **Early**? 2
1.1.3 Why Assess **Orally**? 4
1.1.4 The Early Grade Reading Assessment's (EGRA's) Place Among Assessment Options 5
1.2 Development of the EGRA Instrument 6
1.3 The Instrument in Action 7
1.4 The Original Toolkit and Second Edition 9
1.5 Using the Toolkit 10

2 Ethics of Research and Mandatory Review by Institutional Review Board (IRB) 12

2.1 What Is an IRB? 12
2.2 How Does IRB Approval Apply to EGRA Studies? 13
2.3 Informed Assent and Consent by Participating Individuals 14

3 Purpose and Uses of EGRA 15

3.1 History and Overview 15
3.2 EGRA as a System Diagnostic 16

4 Conceptual Framework and Research Foundations 19

4.1 Summary of Skills Necessary for Successful Reading 19
4.2 Phonological Awareness 20
4.2.1 Description 20

4.2.2 Measures of Phonological Awareness 21
4.3 The Alphabetic Principle, Phonics, and Decoding 21
4.3.1 Description 21
4.3.2 Measures of Alphabet Knowledge and Decoding Skills 23
4.4 Vocabulary and Oral Language 24
4.4.1 Description 24
4.4.2 Measures of Vocabulary 25
4.5 Fluency 25
4.5.1 Description 25
4.5.2 Measures of Fluency 26
4.6 Comprehension 27
4.6.1 Description 27
4.6.2 Measures of Reading Comprehension 27

5 Designing EGRA Studies 29
5.1 Considerations for EGRA Study Design 29
5.2 Design Options, by Research Purpose 30
5.2.1 Snapshot Assessments and Performance Evaluations as Research Designs 30
5.2.2 Impact Evaluations as a Research Design 31

6 EGRA Instrument Design: Adaptation Development and Adaptation Modification 35
6.1 Adaptation Workshop 35
6.1.1 Overview of Workshop Planning Considerations 36
6.1.2 Who Participates? 37
6.1.3 What Materials Are Needed? 38
6.2 Review of the Common Instrument Components 39
6.2.1 Listening Comprehension 41
6.2.2 Letter Identification 43
6.2.3 Nonword Reading 49
6.2.4 Oral Reading Fluency with Comprehension 51
6.2.5 Phonological Awareness – Identification of Initial or Final Sounds; Letter Sound Discrimination 54
6.2.6 Familiar Word Reading 58
6.3 Review of Additional Instrument Components 61
6.3.1 Dictation 61
6.3.2 Phoneme Segmentation 63
6.3.3 Maze and Cloze 66

6.4 Reasons for Exclusion of Other Potential Instrument Components ... 67
6.5 Translation and Other Language Considerations ... 68
6.5.1 Translation vs. Adaptation ... 68
6.5.2 Cross-Language Comparisons: Preparations and Considerations ... 70
6.6 Using Same-Language Instruments Across Multiple Applications: Creation of Equivalent Test Forms ... 72
6.7 Best Practices ... 73

7 Using Electronic Data Collection ... 74
7.1 Cautions and Limitations to Electronic Data Collection ... 75
7.2 Data Collection Software ... 76
7.3 Considerations for Hardware Selection and Purchasing ... 76
7.4 Supplies Needed for Electronic Data Collection and Training ... 77

8 EGRA Assessor Training ... 78
8.1 Recruitment of Training Participants ... 79
8.2 Planning the Training Event ... 81
8.3 Components of Assessor Training ... 82
8.4 Training Methods and Activities ... 82
8.5 School Visits ... 83
8.6 Assessor-Trainee Evaluation Process ... 86
8.7 Measuring Assessors' Accuracy ... 87

9 Field Data Collection: Pilot Test and Full Study ... 91
9.1 Conducting a Pilot EGRA ... 91
9.1.1 Pilot Study Data and Sample Requirements ... 92
9.1.2 Establishing Test Validity and Reliability ... 93
9.1.3 Considerations Regarding the Timing of the Pilot Test ... 96
9.2 Field Data Collection Procedures for the Full Studies ... 97
9.3 Selecting Students ... 99
9.3.1 Student Sampling Option 1: Random Number Table ... 100
9.3.2 Student Sampling Option 2: Interval Sampling ... 100
9.4 End of the Assessment Day: Wrapping Up ... 102
9.5 Uploading Data Collected in the Field ... 103

10 Preparation of EGRA Data ... 105
10.1 Data Cleaning ... 105
10.2 Processing of EGRA Subtasks ... 107

10.2.1 <prefix>_ ... 108
10.2.2 <suffix> ... 108
10.3 Timed Subtasks ... 109
10.4 Untimed Subtasks ... 110
10.5 Statistical Equating ... 111
10.6 Making EGRA Data Publicly Accessible ... 114

11 Data Analysis and Reporting ... 117

11.1 Descriptive Statistics (Non-inferential) ... 117
11.2 Inferential Statistics ... 118
11.3 Types of Regression Analysis ... 119
11.4 Reporting Data Analysis ... 120

12 Using Results to Inform Action ... 122

12.1 A Strategy for Dissemination ... 122
12.1.1 Communicating Results ... 123
12.1.2 Dissemination Approaches ... 125
12.2 Setting Country-Specific Benchmarks ... 129
12.2.1 What Are Benchmarks? ... 131
12.2.2 Criteria for Establishing Benchmarks ... 132
12.2.3 A Process for Setting Benchmarks ... 134
12.3 Cautions and Limitations ... 135

Bibliography ... 137

Annex A: Information About 2015 EGRA Workshops ... 151

Annex B: Sample Size Considerations in Early Grade Reading Assessments ... 154

Annex C: Complex and Cluster Sampling ... 171

Annex D: Sampling for Impact Evaluations ... 173

Annex E: Evaluating the Technical Quality of the EGRA Instrument ... 176

Annex F: Recommendations and Considerations for Cross-Language Comparisons ... 179

Annex G: Comparison of Data Collection Software ... 184

Annex H: Comparison of Paper vs. Electronic EGRA Instructions ... 186

Annex I: Sample Assessor Training Agenda ... 201

Annex J: Data Analysis and Statistical Guidance for Measuring Assessors' Accuracy ... 203

Annex K: Sample Plans for Field-Based Interrater Reliability Testing ... 208

Annex L: Sample Codebook ... 211

Annex M: Recommendations for Equating 213

Annex N: Detailed Technical Recommendations on Public-Use Files 217

Annex O: EGRA Data Analysis 222

Annex P: English Oral Reading Fluency Norms 225

LIST OF EXHIBITS

PAGE

Exhibit 1. Reading trajectories of low and middle readers: Oral reading (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 6

Exhibit 4. Map of EGRA administrations 8

Exhibit 5. Worldwide application of the EGRA instrument: Number of countries, by year 9

Exhibit 6. The continuous cycle of improving student learning 17

Exhibit 7. Differences between EGRA adaptation development and adaptation modification 36

Exhibit 8. Sample agenda: EGRA adaptation development or adaptation modification workshop 39

Exhibit 9. Review of common instrument components 41

Exhibit 10. Sample: Listening comprehension (English) 42

Exhibit 11. Letters in English language: Frequency of use 44

Exhibit 12. Sample: Letter sound identification (Icibemba language, Zambia) 47

Exhibit 13. Sample of letter sound identification subtask with digraphs/diphthongs (Wolof language, Senegal) 48

Exhibit 14. Sample: Nonword reading (Icibemba language, Zambia) 51

Exhibit 15. Sample: Oral reading fluency with comprehension (English) 53

Exhibit 16. Sample: Phonemic awareness – Initial sound identification (English) 55

Exhibit 17. Sample: Phonemic awareness – Letter sound discrimination (Bahasa Indonesia) ... 57

Exhibit 18. Sample: Familiar word reading (Portugese, Timor-Leste) ... 59

Exhibit 19. Sample: Dictation – letter writing (Creole, Haiti) ... 62

Exhibit 20. Sample: Phoneme segmentation (Portugese, Timor Leste) ... 64

Exhibit 21. Sample: Maze (English, Kenya) ... 67

Exhibit 22. Frame from video used for assessment ... 88

Exhibit 23. Sample protocol for monitoring interrater reliability during fieldwork ... 89

Exhibit 24. Differences between EGRA pilot test and full data collection ... 92

Exhibit 25. Determinants of the sampling groups ... 101

Exhibit 26. Data cleaning checklist ... 106

Exhibit 27. EGRA subtask variable nomenclature and names of the timed score variables ... 107

Exhibit 28. Suffix nomenclature for the item and score variables ... 109

Exhibit 29. Sample counterbalanced design ... 113

Exhibit 30. A framework for communication ... 124

Exhibit 31. Overview of potential audiences ... 124

ABBREVIATIONS

ACER	Australian Council for Educational Research
AIR	American Institutes for Research
ASER	Annual Status of Education Report (Pratham)
CFR	US Code of Federal Regulations
CI	confidence interval
CLP	Community Livelihoods Program (Yemen)
clpm	correct letters per minute
clspm	correct letter sounds per minute
cnwpm	correct nonwords per minute
CONFEMEN	Conférence des Ministres de l'Éducation des Pays ayant le Français en Partage
csspm	correct syllable sounds per minute
CTOPP-2	Comprehensive Test of Phonological Processing, Second Edition
CTT	classical test theory
cwpm	correct words per minute
DDL	Development Data Library
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
DID	differences in differences
DIFF	hypothesized difference
EDC	Education Development Center, Inc.
EdData II	Education Data for Decision Making (USAID program)
EFA	Education for All
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
FLAT	Functional Literacy Assessment Tool (World Vision)
GIZ	German aid agency, Deutsche Gesellschaft für Internationale Zusammenarbeit
GPC	grapheme–phoneme correspondence

GPS	global positioning system
ICC	intra-class correlation coefficient
IPA	International Phonetic Alphabet
IRB	Institutional Review Board
IRR	interrater reliability
IRT	item response theory
KNEC	Kenya National Examinations Council
L1, L2	first language, second language
LCD	liquid crystal display
LLECE	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación
LQAS	lot quality assurance sampling
MDES	minimum detectable effect size
MDG	Millennium Development Goal
MoEST	Ministry of Education, Science and Technology (Kenya)
MSI	Management Systems International
NCFL	National Center for Family Literacy
NICHD	US National Institute of Child Health and Human Development
OLS	ordinary least squares
ORF	oral reading fluency
PASEC	Programme d'Analyse des Systèmes Educatifs de la CONFEMEN
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PPS	probability proportional to size
PPVT	Peabody Picture Vocabulary Test
PRIMR	Primary Math and Reading Initiative (Kenya)
PUF	public-use file
QED	quasi-experimental design
RCT	randomized controlled trial
RTI	RTI International (registered trademark and trade name of Research Triangle Institute)

SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SART Ed	Secondary Analysis for Results Tracking project, Education portal
TIMSS	Trends in International Mathematics and Science Study
TOPA-2+	Test of Phonological Awareness, Second Edition Plus
UNDP	United Nations Development Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
URC	University Research Co., LLC
USAID	United States Agency for International Development
YEGRA	Yemen Early Grade Reading Approach

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.*

Diacritic. A mark or sign (such as a grave accent – è, circumflex – ô, or umlaut – ü) that when added to a grapheme indicates a difference in pronunciation of the grapheme.

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.

Census. When all members of the *population* are included in a study (that is, no *sampling* is conducted).

Cluster sampling. A sampling technique whereby the population is divided into groups (or clusters); the clusters are sampled; and then all items within the cluster are evaluated. For instance, a complete list of all primary schools might be used to sample 20 schools, then all grade 3 students in those selected schools would be assessed.

Complex sampling / mixed sampling. A sampling technique similar to *cluster sampling* (see above), but items within the sampled unit are further sampled. For instance, a complete list of all primary schools might be used to sample 20 schools, then 10 grade 3 students would be further sampled (and assessed) within the selected schools.

Confidence interval (CI). A range of values around a value measured from a sample that reveals how precisely the sample value reflects the population value. A larger confidence interval reflects lower precision. For example, if the average age of a sample is 36, then a smaller confidence interval (from 35 to 37) suggests that the sample average age is likely a more precise estimate of the population average age than if the confidence interval were larger (ranging from 34 to 38, for example).

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 gauge 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.

Minimum detectable effect. The smallest treatment effect that can be observed from the data, given a certain sample size.

Nonprobability sample. Any sampling procedure in which samples are selected without the use of probability theory. Examples include *convenience, snowball,* and quota sampling (see entries for italicized terms).

Point estimate. A single value or effect size, derived from the sample data, which provides an estimate of the value or effect size for the *population* (see below) as a whole.

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.

Power analysis. Power analysis can be used to calculate the minimum *sample* size required such that one can be reasonably likely to detect an effect of a given size. Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size.

Precision estimate. When several samples are drawn from a population, a precision estimate is how close *point estimates* (see above) from the different samples are to each other. The closer these point estimates are to each other, the more precise is the estimate.

Probability sample. This is a general term for all samples that are selected in accordance with probability theory, typically involving a random-selection mechanism. Common examples include probability proportional to size (PPS) and simple random sampling.

Propensity score matching. This procedure involves matching observations from a treatment group (treated) and a comparison group (untreated) on the estimated probability of participating in a program (given a range of observed characteristics). The purpose of this approach is to balance the treatment and control groups for analysis, particularly when random assignment was not used for program participation.

Regression discontinuity. A quasi-experimental research design used to estimate treatment effects in a nonexperimental setting by exploiting a cutoff or threshold (upon which an intervention is assigned). For example, if a reading program were to be assigned to students who scored below 50 on an assessment, this approach would focus on the students just below and just above that cut score (based on the assumption that they were similar students but happened to be assigned to the intervention group—or not—due to the virtual randomness of being only a point above or below the threshold).

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), from a population, selected to be in a study.

Sampling frame. The list of all the members of a *population* that can be accessed for inclusion in the *sample* (see entries for italicized terms). The sample is drawn from the sampling frame.

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).

Snowball sample. Nonprobability sample in which initial sample participants are used to provide information necessary to locate additional sample participants.

Statistical power. The probability that a study will find a treatment effect given that there is a treatment effect to be detected.

Statistical significance. The likelihood that a treatment effect found in a study is not a result of chance. A higher statistical significance indicates a higher likelihood that the observed treatment effect is not the result of chance.

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

Methodological Terms

Assessor drift. The propensity of assessors to change what they consider an acceptable, "correct" answer over time.

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

Content validity. Term used to indicate the degree to which 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* (see below) 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* (see below) following the treatment. Frequently selected based on similarity of certain characteristics with the treatment group (also called "matched comparison group").

Counterfactual. A measure of what would have happened to a treatment group in the absence of treatment. Because the true counterfactual is unknowable, a variety of statistical methods are used to construct a counterfactual group that represents what would likely have happened to the treatment group in absence of treatment. The treatment group is then compared against this counterfactual to obtain an estimate of the treatment effect.

Cross-sectional design. A research design in which data are collected from the same *sample* (see Statistical Terms) only once. Data from these designs can be compared with other data that are independently drawn from the same population at different times. Example: A trend analysis of grade 3 students in 2016 compared to grade 3 students in 2017.

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 of one's opinion to which a test overall is viewed as covering the concepts its designers intended to measure.

Longitudinal (panel) design. A study design in which the same *sampling units* (see Statistical Terms) are tracked over a period of time and data are collected from the same sampling units repeatedly.

Snapshot design. A type of *cross-sectional* study (see above) for which data are collected only once and no comparisons are made over time.

Stratification. The process of separating members within a population into subgroups before they are sampled. Stratification is often used in sampling to ensure adequate sample size of each subgroup within the population.

Treatment group. Subjects who receive an intervention. Also called intervention group.

1 INTRODUCTION

1.1 Why Do We Need Early Grade Reading Assessments?

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 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 fell 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, 2015a). 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); 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 measures 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 and skill 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 will 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**?

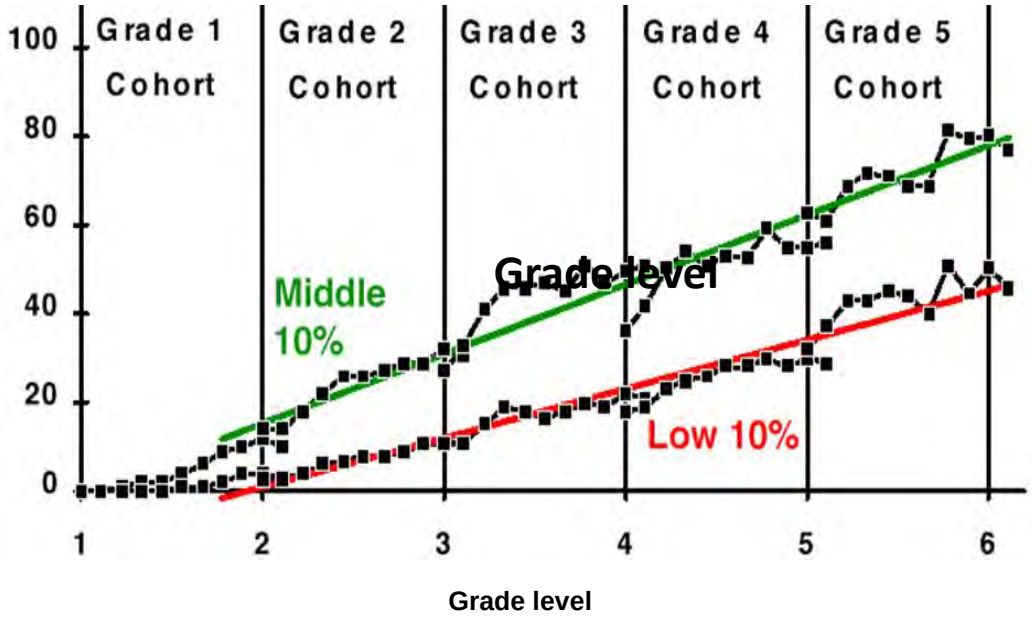
Basic literacy is the foundation children need to be successful in all other areas of education. Children first need to "learn to read" so that they can "read to learn." That is, as children pass through the grade levels, more and more academic content is transmitted to them through text, and their ability 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 and reading comprehension. 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 to eventually drop out of school. That is, if strong foundational skills are not acquired early on, gaps in learning outcomes (between students who have mastered foundational reading skills and those who have not) grow larger over time (see **Exhibit 1** as well as Adolf, Catts, & Lee, 2010; Daniel et al., 2006; Darney, Reinke, Herman, Stormont, & Ialongo, 2013; Scanlon, Gelzheiser, Vellutino, Schatschneider, & Sweeney, 2008; Torgesen, 2002). The common metaphor of "the rich get richer 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: Oral reading (measured in correct words per minute)



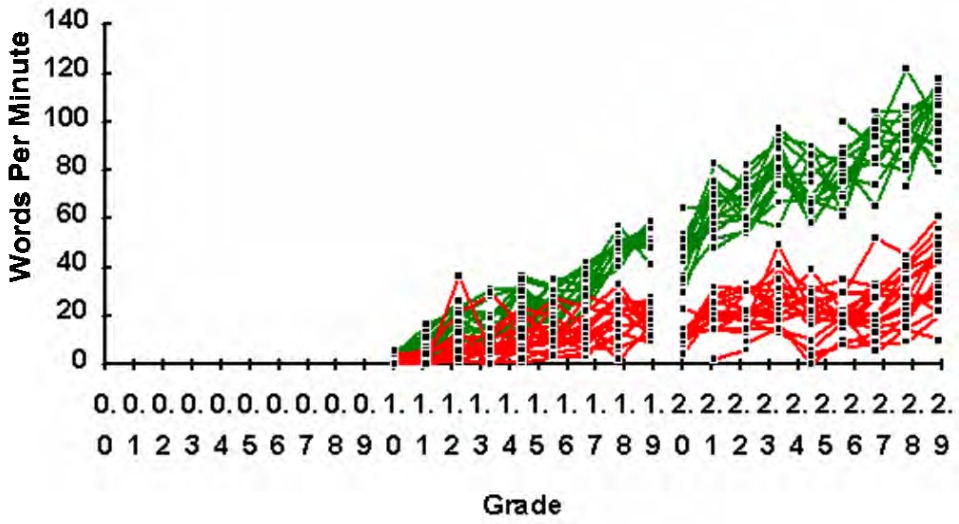
Source: Good, Simmons, & Smith, 1998, Figure 1.

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 children 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 reading 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 test will not be able to accurately measure what children know. In technical terms, the results will suffer from a floor effect, with a high number of students attaining 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. 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 students not have the knowledge to answer the questions, or were they just unable to read the questions?

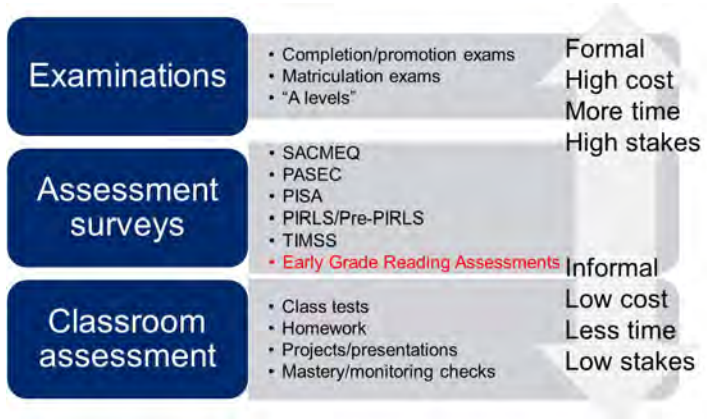
Reading fluency and comprehension are higher-order skills in the reading acquisition process, and they build upon several lower-order, foundational skills such as phonological awareness, alphabet knowledge, decoding, vocabulary, etc., which can be detected through an oral assessment. An oral assessment therefore can give us more information about what students actually do know and where they are in the reading acquisition process early on. Oral assessments can also help detect 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 The Early Grade Reading Assessment's (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 tests typically are longer, more formal, standardized 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 informing and improving teaching and learning (Kanjee, 2009). These less formal assessments often come in the form of classroom tests, homework assignments, and projects/presentations. By design, classroom assessments are intended to be

inexpensive, to take less time, and to involve lower stakes (particularly when compared with examinations).

Exhibit 3. Different types of assessments: A continuum



Source: Adapted from Kanjee (2009).

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[2] (PASEC), the Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), and the Trends in International Mathematics and Science Study (TIMSS). Because the tests associated with these programs are intended to measure trends in 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 may limit the usefulness 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][3] assessment) have been developed 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 reading outcomes (Abadzi, 2006; Center for Global

[2] CONFEMEN: Conférence des Ministres de l'Éducation des Pays ayant le Français en Partage.

[3] Functional Literacy Assessment Tool developed and used by World Vision: http://www.wvi.org/development/publication/functional-literacy-assessment-tool-flat

Development, 2006; Chabbott, 2006; World Bank: Independent Evaluation Group, 2006).

To respond to this demand and the need for a low-cost and effective way to measure early reading acquisition, 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 an accurate, 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 process 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.

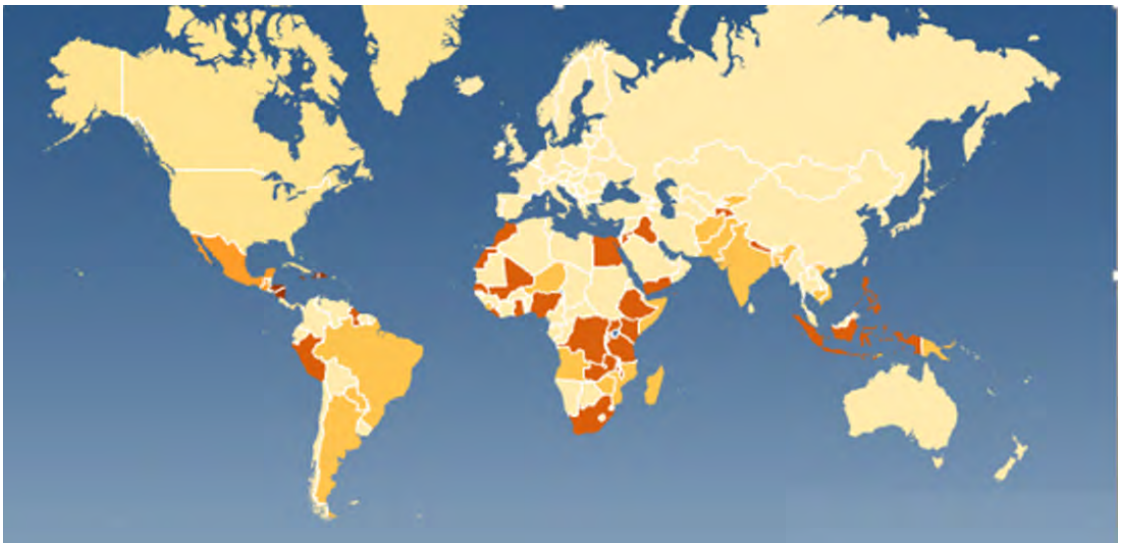
EGRA is open source and readily available to support a higher level and wider dissemination of knowledge on reading and learning outcomes. The purpose behind this decision was to ensure that both technical and nontechnical audiences could acquire accurate, timely, and accessible data regarding early literacy and numeracy in 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 (see **Exhibit 4**).

Exhibit 4. Map of EGRA administrations

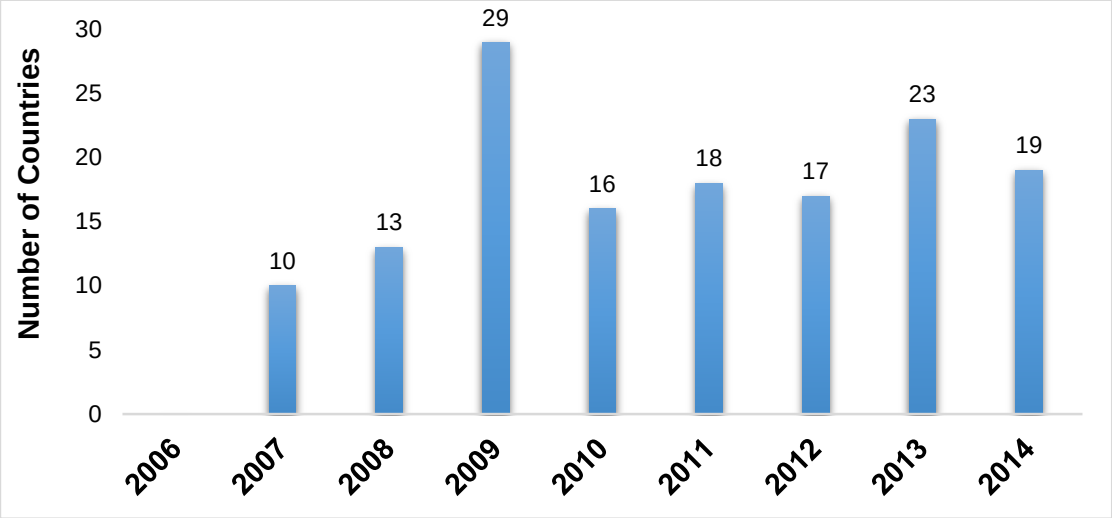


Source: RTI International for the EdData II project website, https://www.eddataglobal.org/countries/index.cfm

As of September 2015, nearly 10 years after the initial development of EGRA, the tool had been used by over 30 organizations in more than 70 countries. The early grade reading approach also shifted to focus on mother-tongue instruction,[4] 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.**

[4] An individual's first language (L1) is the one he or she speaks best, often referred to as a *mother tongue* or *home language.* A second language (L2) is a language that someone learns in addition to his or her first language.

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



Data source: RTI International, 2015.

1.4 The Original Toolkit and Second Edition

In the interest of consolidating diverse experiences and developing a reasonably standardized approach to assessing children's early reading acquisition, in 2009 the World Bank requested that RTI develop a "toolkit," or user manual, which would serve as a guide for countries beginning to work with EGRA in such areas as local adaptation of the instrument, fieldwork, and analysis of results.

As use of the EGRA instrument became more widespread among implementing organizations and countries, largely due to the open-source nature of the assessments, USAID commissioned RTI to revise and update the toolkit. Through an EdData II task order, Measurement and Research Support to Education Strategy Goal 1, RTI spearheaded the development of this second edition of the EGRA toolkit. This revised version reflects progress made since the original version of the toolkit toward improving the quality of data being used to advance the agenda of the Sustainable Development Goals.

The toolkit revision process began in December 2014. RTI started by compiling EGRA-related experiences and information from across all its EdData II task orders and other USAID-funded studies. Resources encompassed project reports, published research, and anecdotal information from international education researchers. In addition to data analysis reports on various countries' results, it included materials to reflect best practices, lessons learned from various EGRA implementations, and new

technological advances in regard to EGRA planning and implementation. The information was reviewed and condensed into presentations and handout materials.

The presentations and handout materials were then used for an EGRA workshop titled “Designing and Implementing Early Grade Reading Assessments: Understanding the Basics,” which was hosted by the Global Reading Network as a global workshop and webinar in March 2015.

Shortly after the March workshop on EGRA basics, a workshop to further improve the quality of EGRA data was held in May 2015. Again structured as a workshop and a webinar, it was hosted by the Global Reading Network and funded by USAID. Experts from various organizations presented on EGRA design, administration, analysis, and reporting. The presentations were followed by facilitated discussions between the workshop participants and the panelists. Additional ideas, as well as the ideas presented by the panelists, were debated and discussed.

The aim of these workshops was twofold: to expand the use of EGRA by presenting detailed training on how to conduct an EGRA study, as well as to improve the quality of EGRA data by opening up the evaluation and analysis processes. For more details regarding these workshops, the format, and the participants, see **Annex A.**

The next step in the toolkit-update process came after the conclusion of the two workshops. Working groups were formed with technical experts. The working groups were tasked with discussing and agreeing upon a final set of recommendations regarding the panel session topics presented at the May 2015 workshop. This collaborative effort included ideas and consensus from multiple implementing organizations, in an effort to present well-defined methodologies for planning, implementing, and analyzing EGRA data.

This updated toolkit is a product of these workshops and technical working groups. It represents the collaborative input of multiple organizations and individuals across the international development and education fields.

1.5 Using the Toolkit

This toolkit is intended for use by Ministry or Department of Education staff, donor staff, practitioners, and professionals in the field of education development. The document, in 12 sections, seeks to summarize a large body of research in an accessible manner. The procedures described in this toolkit are to be used in all USAID-funded administrations of EGRA and, it is hoped, in all other EGRA administrations as well.

The toolkit is not intended to be a comprehensive review of all reading research. Even with the new contributions from other organizations and individuals, in the interest of brevity, the toolkit does not cover all aspects of and alternatives to reading assessment. It should also be noted that it is not an “off-the-shelf” assessment guide—each new country application requires a review of vocabulary and development of context-appropriate reading passages. Those seeking specific guidance on planning and implementing EGRA should reference the *Guidance Notes for Planning and Implementing EGRA* (RTI International & International Rescue Committee, 2011).

Following this introduction, Section 2 covers the topic of the protection of human subjects in research. Section 3 is an overview of the purposes and uses of EGRA. Section 4 addresses the conceptual framework and research foundations (the theoretical underpinnings of the assessment). Section 5 discusses options for study design. Section 6 discusses preparatory steps to administration of the assessment, including the design workshop for construction of the EGRA instrument. Section 7 is an overview of electronic data collection. Section 8 provides information and procedures regarding assessor training. Section 9 advises on field data collection for both pilot testing and full EGRA studies. Section 10 discusses appropriate protocols for cleaning and preparing survey data. Section 11 is an overview of analyses to be conducted. Finally, Section 12 provides guidance on interpretation of results, establishment of benchmarks, and some summary implications for policy dialogue related to improving instruction and reporting results to schools.

A set of annexes expands on points made in the text with examples and illustrations, technical details, and statistical guidance.

2 ETHICS OF RESEARCH AND MANDATORY REVIEW BY INSTITUTIONAL REVIEW BOARD (IRB)

Research institutions receiving federal funds must follow US federal regulations for conducting ethical research and the United Nations' Fundamental Principle of Official Statistics, which states: "Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes."[5] All US organizations that conduct research involving human subjects must consult an Institutional Review Board (IRB) in advance of a survey (22 CFR 225). The Nuremberg Military Tribunals led to the formulation of Institutional Review Boards and human subject protections, and US regulations regarding the protection of human subjects began in 1974.

All US organizations that conduct research involving human subjects must consult an Institutional Review Board in advance of a survey.

2.1 What Is an IRB?

IRBs use the set of basic principles outlined in the "Belmont Report," issued in the United States by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1978), to guide their review of proposed research protocols. The Belmont Report outlines three basic principles:

- **Respect for persons.** Potential research subjects must be treated as autonomous agents, who have the capacity to consider alternatives, make choices, and act without undue influence or interference from others.

[5] Federal policy for protection of human subjects is required by the US Code of Federal Regulations, 22 CFR Part 225.

- **Beneficence.** The two basic principles of beneficence are: (1) do no harm, and (2) protect from harm by maximizing possible benefits and minimizing possible harm.
- **Justice.** This ethical principle requires fairness in the distribution of the burdens and benefits of research.

Additional guidelines for evaluating human subjects were produced by the US Food and Drug Administration and the US Department of Health and Human Services in 1981. The criteria the two agencies delineated are as follows:

1. The protocol must be evaluated to see if it is scientifically sound and worthwhile;
2. Risks must be minimized to the extent possible;
3. Subjects must be selected in an equitable manner;
4. Informed consent is required;
5. Privacy and confidentiality must be protected;
6. The study must be adequately monitored.

Research is defined as "a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge."

–US Code of Federal Regulations, 22 CFR 225

2.2 How Does IRB Approval Apply to EGRA Studies?

As mentioned earlier, all organizations that receive support via US federal funds or that are otherwise subject to regulation by any federal department or agency and are conducting research which involves human subjects are required to consult an IRB and receive IRB approval before conducting the research. Research is defined as "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge." Institutional Review Boards are responsible for reviewing research projects that involve human subjects and determining the degree of risk that subjects may experience as a result of participating in the research. Research activities are approved or denied by the IRB accordingly based on the thorough review of survey protocols and circumstances in which the research is being conducted.

Research that incorporates educational tests is often found to be exempt from IRB requirements (based on the philosophy that the tests administered do not differ

greatly from what children experience in their natural school environments). Nonetheless, only an IRB is able to decide whether an EGRA-related study receives exempt status. If an IRB decides to give an exemption to an overall study, preapproval of survey questions still may be required specifically, if information collected during the survey could put students or teachers at risk.

In the case of EGRA and related assessments of young children, every country that authorizes an EGRA study also must be given an opportunity for its own ethical body to review the study terms and issue its approval to go forward, or request any modifications needed to warrant such approval (22 CFR 225).

2.3 Informed Assent and Consent by Participating Individuals

The EGRA template and supporting instruments always contain a section at the beginning instructing assessors on how to request consent (from adults) or assent (from children) of those selected to participate. Prior to administering the EGRA to children, assessors describe the objectives of the study and inform students that the assessment is anonymous, will not affect their grade in school, and will be used to make improvements in how children in their country learn to read. Each child has the option of orally assenting to be assessed or declining to participate, without consequences of any kind. If school principal or teacher surveys are conducted as part of the study, a similar consent process—in writing rather than orally—is completed.

Although this assent/consent process may be unfamiliar to host-country counterparts, the process is often welcomed by students and teachers, who report feeling empowered at being given the option to participate. Experience across multiple EGRA implementations to date has shown that few students and teachers decline to participate. If an eligible participant declines to participate, another respondent is randomly selected. For additional information on IRBs and ethical research with human subjects, including children, please see the website of the US Department of Health and Human Services, Office for Human Research Protections, http://www.hhs.gov/ohrp.

3 PURPOSE AND USES OF EGRA

3.1 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 3.2) 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.[6] 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

[6] 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.

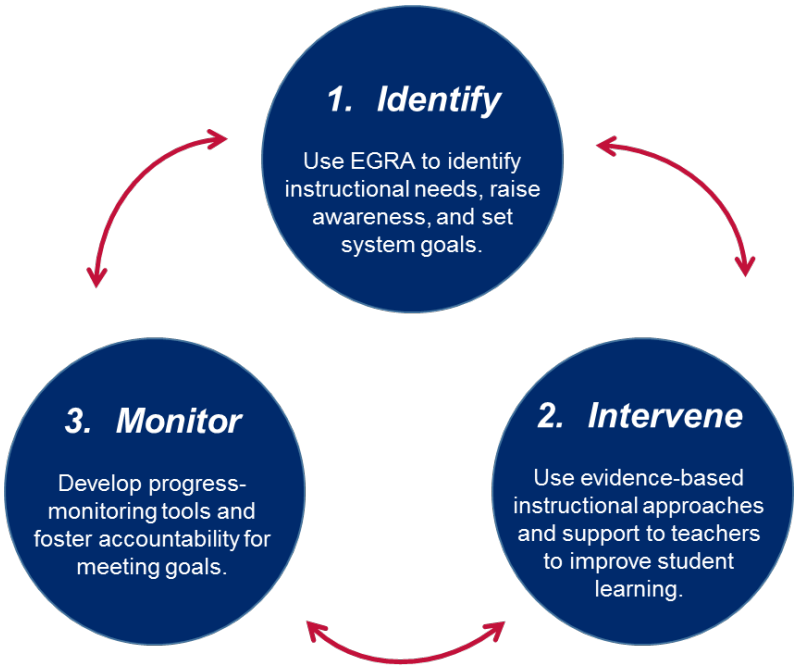
promotion or to evaluate individual teachers. EGRA is designed to complement, rather than replace, existing curriculum-based pencil-and-paper assessments. EGRA is made up of a set of subtasks that measure foundational skills that have been found to be predictive of later reading success. However, due 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 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.

3.2 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 (Dubeck & Gove, 2015). General benchmarking and creation of goals for future applications (see Section 12.2) 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 teacher training and professional development 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. The basis for this approach is the understanding that the ways in which students are learning as a whole is a direct reflection of the instruction they are receiving. 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 must relate to existing performance expectations and benchmarks, (2) the assessment must correlate with later desired skills, and (3) it must be possible to modify or improve upon the skills through adjusted 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 undertaken in at least 12 countries to draft national or regional benchmarks using EGRA data. Details are discussed in Section 12.2.

In addition, the EGRA tasks were developed intentionally to be predictive of later reading achievement, and numerous administrations of EGRA in multiple countries and languages have 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 adjustments to instruction. EGRA is valuable as a diagnostic tool precisely because it includes measures of those skills that can be improved through instruction.

4 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.

4.1 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 complex skill or a composite behavior that is built when a wide array of subskills are mastered and used 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 comprehension. The EGRA subtasks (refer to Section 6) 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) or ceiling effects (almost all children receive high scores), depending on where the children are in their development.

4.2 Phonological Awareness

4.2.1 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/[7], /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—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" (see *diacritic* in glossary). 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

[7] Phonemes are traditionally written between slashes in the International Phonetic Alphabet (IPA). 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.

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).

4.2.2 Measures of Phonological Awareness

EGRA instruments typically include one or two measures of phonological awareness, usually at the phonemic level (i.e., phonemic awareness). These include **initial sound identification** subtasks, in which children are presented with a word orally and asked to isolate and pronounce only the first sound of the word; or an **initial sound discrimination** task, in which the children are presented with three words and asked to pick out the word with a beginning sound that differs from the other two.

An optional (i.e., not core) **phoneme (or syllable) segmentation** subtask, in which the children are presented with a word orally and asked to pronounce its component phonemes (or syllables) individually, has also been used. Because segmentation is a more complex skill (Linan-Thompson & Vaughn, 2007), this subtask has produced floor effects in many low-performing contexts. It can be an appropriate alternative, however, when the initial sound identification subtask produces a ceiling effect.

4.3 The Alphabetic Principle, Phonics, and Decoding

4.3.1 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-Dağli, 2011), for both native and nonnative speakers of a language (Chiappe, Siegel, & Wade-Woolley, 2002; Manis, Lindsey, & Bailey, 2004; Marsick & Watkins, 2001; McBride-Chang & Ho, 2005). 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.

4.3.2 Measures of Alphabet Knowledge and Decoding Skills

EGRA assesses children’s alphabet knowledge in several ways, beginning with the **letter identification** subtask, a component of the core EGRA. The letter identification subtask tests children’s ability to recognize the graphemic features of each letter and accurately map it to its corresponding name or sound. Either or both letter identification subtasks can be selected, depending on what is appropriate for a given context: **letter name identification** or **letter sound identification.** In both options, 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.

Originally, letter name identification was the more widely used measure of alphabet knowledge within EGRA assessments, and it was shown to be a strong predictor of later reading achievement in English. However, over time, letter sound identification has become the more frequent option, as letter sound knowledge is more directly linked to the children’s ability to decode words, especially in transparent orthographies (Ehri, 1998).

EGRA developers may choose to incorporate a **syllable identification** measure in addition to letter names or sounds. This task has been used in contexts where the language has primarily open (i.e., vowel-final) syllables and/or where the reading pedagogy in that language stresses syllabic combinations.

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.

The **familiar word reading subtask** is similar in format to the nonword reading subtask except that it presents a list of words that children are expected to be able to read at their grade level and will have likely encountered before. Again, according to the dual-route model, children are more likely to process familiar words—if they are indeed familiar, and especially if they have irregular spellings—directly by the lexical route. That is, they might recognize the words instantly, rather than attempting to decode them sound by sound.

Finally, **dictation** asks students to listen to letter sounds, words, and/or a short sentence and then write them down. The subtask measures students' alphabet knowledge and ability to hear and distinguish the individual letter sounds in isolation or in words and to spell (encode) words correctly. If a sentence is presented, the task may also measure their ability to use correct sentence-writing conventions such as capital letters and punctuation. This subtask has proven challenging to score in a standardized way in some contexts. It is no longer part of the core instrument but has been used in some countries that have found it appropriate.

4.4 Vocabulary and Oral Language

4.4.1 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).

4.4.2 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 4.6) in order to tease out whether comprehension difficulties stem primarily from low reading skills or from low overall language comprehension.

4.5 Fluency

4.5.1 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). Automaticity may also be critical due to the constraints of our short-term working memory. 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, Cummiskey, & 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 to indicate reading proficiency.

4.5.2 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.

4.6 Comprehension

4.6.1 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).

4.6.2 Measures of Reading Comprehension

EGRA measures reading comprehension through the **reading comprehension** subtask, based on the passage 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. An optional **maze** or **cloze** subtask, asking children to identify a word among several choices that would complete a sentence using the correct part of speech, has the potential to further assess comprehension but has not been commonly used in EGRA administrations to date.

EGRA was designed to be a large-scale, standardized measure. Its design reflects the fact that research has not yet produced a proven means to consistently and thoroughly test the higher-level and more nuanced comprehension skills in a

standardized way that could be accepted as valid and reliable. Some existing alternative measures—such as a story retell activity that would ask students to tell what they remembered from the story, and would allow a teacher to probe for deeper comprehension as the student talked—have the disadvantage of inherently requiring open-ended responses and subjective response scoring. However, additional options are under consideration, and it is hoped that measurement of comprehension will continue to improve, as this skill is one of the most important measures of reading success.

5 DESIGNING EGRA STUDIES

5.1 Considerations for EGRA Study Design

This section establishes research designs and principles for EGRA-based assessments and evaluations. As with any research, the purpose of the assessment or evaluation, and the research questions that drive it, guide and tailor the design. All of these questions and decisions are considered in close collaboration with donors and host government counterparts to ensure both the feasibility and appropriateness of the design decision.

REQUISITES FOR SURVEY SAMPLES, FOR ALL RESEARCH DESIGNS

EGRA research designs and sample designs are closely linked and interdependent; however, all EGRA survey samples, regardless of the research design, include the following in the data, analysis, and reporting:

- An explicit **definition of the population of interest** with all of the exclusions well documented prior to sampling;
- Properly calculated **sampling weights** (used to weight the sample to the population from which it was drawn);
- Properly set up **complex survey analysis** (to account for the sample methodology and cluster effects);
- **Statistical software** that can properly analyze the complex sample (e.g., SPSS, Stata, SAS);
- **Inferential analyses** using the proper complex survey analysis setup.

The first step in determining the design is to ask a question that must not be overlooked in the rush to measure—**What is the purpose of the EGRA study?** EGRA studies typically fall into one of three categories:

1. **Snapshot assessment** – to obtain a diagnostic of student performance at a single time point
2. **Performance evaluation** – to evaluate whether changes occurred in learners' performance over a period of time, based on initial and follow-up assessments
3. **Impact evaluation** – to evaluate the impact of a program or intervention on learners' performance over a period of time, based on a comparison of treatment and control groups.

Section 5.2 elaborates on these purposes and the most appropriate study designs associated with each type of study. Each design option is described, with accompanying annexes presenting detailed sampling information associated with the designs.

5.2 Design Options, by Research Purpose

5.2.1 Snapshot Assessments and Performance Evaluations as Research Designs

As noted above, snapshot assessments seek to provide a view of a particular indicator or variable, such as oral reading fluency, at one time point; performance evaluations do the same at multiple points. Neither snapshot assessments nor performance evaluations allow for attribution of results to a specific intervention.

SAMPLING CONSIDERATIONS: SNAPSHOT ASSESSMENTS AND PERFORMANCE EVALUATIONS

Most EGRA snapshot assessments seek to estimate reading ability for a defined population. These types of assessments typically use complex/cluster sampling (sampling is described in **Annexes B and C**).

By contrast, sampling for performance evaluations is based on the evaluation questions being answered as well as the level of resources available to answer that question. For instance, if the evaluation question asks, "How are beneficiary students performing in relation to nationwide benchmarks?" the performance evaluation will likely need to involve a simple random sample or complex/cluster sampling to ensure a sample that is representative of the entire beneficiary population. But if resources are scarce, the performance evaluation may, instead, require taking a smaller, non-representative sample and afterward explaining the extreme limitations of the data and findings used for statistical inference regarding the full population. This could be done for internal review, but should not be done if the study is trying to generalize to the population of interest.

5.2.2 Impact Evaluations as a Research Design

An impact evaluation differs from a performance evaluation in that impact evaluation attempts to isolate the impact of an intervention on a key outcome from other influences by comparing outcomes for a group receiving treatment (or multiple groups receiving different treatment arms) to a group serving as a control.

In other words, an impact evaluation, using a *counterfactual* (see glossary), reveals how much change in a particular outcome can be confidently attributed to a particular program. A growing number of impact evaluations seek to understand the impact of various early grade reading interventions on EGRA scores. There are two main types of impact evaluation designs—*experimental* and *quasi-experimental* designs—as described in the text box below. Detailed information on sampling for impact evaluations appears in **Annex D.**

TWO TYPES OF IMPACT EVALUATIONS

Experimental Designs

Experimental designs (sometimes referred to as randomized controlled trials, or RCTs) must begin before an intervention starts. They require baseline data and randomization of intervention participants (or schools, zones, or some other type of unit) into a beneficiary group and a nonbeneficiary (or comparison) group. All individuals (or units) must have an equal likelihood of assignment to the beneficiary or comparison group, and sample sizes of each must be large enough to allow for comparison between the two groups with a reasonable *minimum detectable effect size* (MDES; see Annex D).

Quasi-Experimental Designs (QEDs)

Quasi-experimental designs usually begin before an intervention starts, but they do not necessarily have to, as long as baseline data exist for a beneficiary group and possible comparison group. In QEDs, intervention participants (or schools, zones, or some other type of unit) are not randomized into beneficiary and nonbeneficiary groups; instead, participants self-select into an intervention, or the implementer selects the beneficiaries using some sort of selection criteria. Both types of selection options just mentioned typically result in some selection bias, and QEDs attempt to minimize or control for that selection bias through a variety of statistical techniques.

While QEDs still allow for attribution of results, they rely on statistical assumptions that may not always hold true and are, thus, less rigorous and less reliable than experimental designs (assuming both are done well). Some common types of QEDs used for EGRA impact evaluations are *regression discontinuity designs* and *propensity score matching designs* (see glossary).

Levels of Assignment for Impact Evaluations

If it has been decided to use an impact evaluation to measure results, the level of assignment then must be determined as well as whether the study will be *longitudinal* (following the same students over time), *semi-longitudinal* (following the same teachers or schools over time), or *cross-sectional* (resampling different schools and students at each data collection time point). These decisions again depend on the purpose of the study, but they also depend on how the intervention (program, project, or activity) being assessed will be implemented. There are multiple levels at which an intervention may provide benefits:

- **District, zone, or administrative unit level** – e.g., provide teacher training for all teachers in a specific district
- **Community level** – e.g., carry out community outreach programming to get communities more involved in schools or start a community reading center
- **School level** – e.g., supply books and materials or other benefits directly to schools, targeting some schools in an administrative unit but not others
- **Student level** – e.g., give some students within a school—but not others—scholarships or conditional cash transfers

Interventions may provide benefits at multiple levels. As such, it is important when determining the level of assignment for beneficiary and comparison groups to make assignments based on the highest level of intervention planned by the implementer. For instance, if a program plans to offer teacher training at the "district level" as well as books or materials at the "school level," then "district level" will be assigned to both beneficiary and comparison groups. For this reason, it is absolutely critical that evaluation teams work very closely alongside implementers to plan the evaluation design as the implementer is designing the intervention.

Study Design Options for Impact Evaluations

Next, it must be determined whether to track students longitudinally or to use a semi-longitudinal or cross-sectional design. The following information is required to help make an appropriate determination:

1. **What is the evaluation purpose and what are the evaluation questions?**

   Longitudinal design. If the research question requires an understanding of specific changes for every student receiving the intervention, using a longitudinal design will allow researchers to definitively detect any changes that occurred in each student participating in the evaluation. However, researchers will not be able to attribute those findings to the general population receiving the

intervention. This type of design is best for pilot studies, and for continuing internal evaluations of an intervention.

Semi-longitudinal design. If the goal of the research question is to investigate intricate changes that are occurring in specific schools, then a semi-longitudinal study is ideal, whereby the same schools are visited but a random sample of students is drawn within the same schools. This type of study will allow researchers to see any changes in the specific schools being assessed but will not allow researchers to generalize their findings to the larger population of schools within the study. This type of design is best for pilot studies, and for continuing internal evaluations of an intervention.[8]

Cross-sectional design. If the purpose of the research question is to assess how a *population* (see glossary) of schools and a population of students within those schools are changing due to an intervention, then a series of cross-sectional samples is needed, whereby completely different samples of schools and students are drawn each time data are collected. The schools and students are sampled from within the population of schools and students receiving the intervention. This design will not allow researchers to determinately detect the exact changes that occurred within specific schools or students (because they are different every time). But the design will allow researchers to generalize changes that can be attributed to the population as a whole. This type of study is best for external evaluation of an intervention study, whereby researchers are trying to better understand the impact of the intervention on the whole population that received the intervention, rather than the impact that the intervention has on specific schools or students.

2. **How easy would it be to track the same students, teachers, schools, etc.?** For instance, if the team is evaluating a program in a country where either every student is assigned a student identification number that remains with him/her even when he/she moves, or households must register with the government for tax or census purposes, it might not be very difficult to track individual students. If this is not the case, though, and if communities tend to be very mobile and dropout rates tend to be very high, tracking the same students over time can be very resource intensive.

3. **What levels of resources are available to track students, teachers, schools, etc.?** Even if tracking is relatively easy in a country, it is still usually more costly and resource intensive than taking a snapshot because some students, for

[8] Longitudinal and semi-longitudinal designs can also be useful for non-intervention studies, as the designs allow for researchers to follow changes of a single unit (e.g., a student) over time in cases in which no intervention is introduced.

instance, may be impossible to track, and as such, oversampling at baseline is usually a necessity.

4. **What level of rigor and precision is needed in the results?** If a donor or implementer needs precise results about dropouts, for instance, a longitudinal study may be necessary. However, if approximate dropout rates as reported by teachers or schools will serve the purpose, a cross-sectional study may suffice.

6 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.

6.1 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 for learning expectations or standards.[9] 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 9.

[9] 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).

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 **Annex E** and Section 9.1.2 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
    - 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.
- Review the procedures for informed consent (adults) or assent (children) and discuss the ethics of research and working with human subjects, especially children.

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

6.1.1 Overview of Workshop Planning Considerations

Whether designing a country-specific EGRA instrument from the beginning (development) or from an existing model (modification), the study 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 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 to gain a deeper understanding of the instrument and serves as a rough test of the items to identify any obvious changes that may be needed (such as revisions to ambiguous answer choices or overly difficult vocabulary).
- 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. Expert panels and an IRB must review the directions to ensure that they are ethically sound and exact for each section. 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 they can be finalized.

6.1.2 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 to 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 number of participants.

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.

6.1.3 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.	*Break*				
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.	*Lunch*				
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.	*Break*				
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:	*Understanding of EGRA purpose and content*	*Oral reading passages and questions developed*	*Listening comprehension passages and stories developed; Pupil Questionnaire Developed*	*Additional subtasks/questionnaires developed*	*Instruments finalized*

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.

6.2 Review of the Common 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 common instrument for letter-based orthographies contains six subtasks, of which four are “core” subtasks. The four are:

1. Listening comprehension
2. Letter identification[10]
3. Nonword reading
4. Oral reading fluency with comprehension

Each of these core components (see also **Exhibit 9**) has been piloted in dozens of languages across the globe.

Two additional common subtasks that are most often included are phonemic awareness and familiar word reading. Phonemic awareness (a subset of phonological awareness) can be assessed by various measures, depending on what may be appropriate for a specific context.[11]

Comments from practitioners and local counterparts have included requests to reduce the number of skills tested in the EGRA. As stated above, one of the goals of the instrument is to assess a reasonably full battery of foundational reading skills to be able to identify which areas need additional instruction. If EGRA were to test only oral reading fluency, many low-income countries likely would see considerable floor effects. Maintaining the number of subtasks around six allows even countries where student reading performance is very weak to determine progress in at least some reading-related skills.

It is also 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 NICHD, 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.

[10] The letter identification subtask can consist of letter sounds or letter names. Although the letter sounds subtask is more commonly used, letter names may be more appropriate depending on the specific country and language where the instrument is being administered. Syllable-based languages may incorporate syllable naming and syllable segmentation subtasks rather than letter naming and letter sounds.

[11] Optional subtasks such as dictation, maze, and cloze are occasionally used in addition to the common subtasks. See the discussion about these additional subtasks in Chapter 1 of Gove and Wetterberg (2011), as well as Section 6.3 below, "Review of Additional Instrument Components."

Exhibit 9. Review of common instrument components

Component	Early reading skill	Skill demonstrated by students' ability to:
1. Listening comprehension	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 names and/or letter sounds	Alphabet knowledge	Provide the name and/or 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 fluency with comprehension	Oral reading fluency	Read a text with accuracy, with little effort, and at a sufficient rate
	Reading comprehension	Respond correctly to different types of questions, including literal and inferential questions about the text they have read
5. Initial or final sound identification, or letter sound discrimination, or phoneme segmentation, identification of onset/rime sounds	Phonological awareness	Identify/differentiate the onset/rime sounds of words or the initial or final sounds of words, or segment words into phonemes by having the assessor and then the student read the phonemes aloud
6. Familiar word reading	Word recognition	Read words which are randomly ordered and drawn from a list of frequent words

6.2.1 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. 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 need to be 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 questions will be literal ones 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 is a sample of the listening comprehension subtask.

Exhibit 10. Sample: Listening comprehension (English)

Sub-test 5. LISTENING COMPREHENSION	📖X			⏱ X
Merebɛkenkan abasɛm tiawa bi baako pɛ dennennen akyerɛ wo, na mabisabisa wo nsɛm kakra afa ho.Mesrɛ wo tie no yiye na bua nsɛmmisa no sɛnea wubetumi biara.Wubetumi de kasa biara a wopɛ ayiyi nsɛmmisa no ano.Metumi afi ase? Yemfi ase. 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.				Remove the pupil stimuli booklet from the child’s view. Do not allow the child to look at the passage or the questions. If a child says “I don’t know,” mark as incorrect.
(✓) 1 = Correct (✓) 0 = Incorrect (✓) . = No response.				
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.				
Why was Issa sad? [he lost his sheep; he could not go to look for his sheep]	1	0	.	
Who went to look for the sheep? [Grandfather]	1	0	.	
Why is Issa smiling now? [Grandfather returned with his sheep; his sheep are back; Grandfather found the sheep]	1	0	.	

Mo! Woayɛ ade. Yɛnkɔ ɔfa a edi so no so. Good effort! Let’s go on to the next section.

6.2.2 Letter Identification

Knowledge of how letters correspond to names or sounds is another critical skill children must master to become successful readers. Letter–sound correspondences are typically taught through phonics-based approaches. Letter identification knowledge is a fairly common assessment approach and is used in several early reading assessments, including the Preschool Comprehensive Test of Phonological and Print Processing (Lonigan, Wagner, Torgesen, & Rashotte, 2002). The assessment can include one or both of the options below. 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. At the same time, research, especially from less transparent orthographies like English, has shown that knowing the names of the letters is also highly predictive of later reading achievement in those languages.

First Approach: Letter Sound Identification

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.[12] The number of times a letter is repeated is based on the frequency with which the letter occurs in the language in question (as an example, see the frequency table for English in **Exhibit 11**). 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 was 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). These tables are available for Spanish, French, and other international alphabetic languages.[13] Test developers constructing instruments in other languages sample 20–30 pages of a grade-appropriate textbook or supplementary reading

[12] 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

[13] Letter-frequency rates for French, German, Spanish, Portuguese, and others are available from University of California at Los Angeles Statistics Online Computational Resource, http://wiki.stat.ucla.edu/socr/index.php/SOCR_LetterFrequencyData#SOCR_Data (accessed September 18, 2015).

material and analyze the frequency of letters electronically to develop similar letter frequency tables.

Exhibit 11. Letters in English language: Frequency of use

E	12.02%	C	2.70%	Y	2.11%
A	8.12%	U	2.88%	W	2.09%
R	6.02%	D	4.32%	K	0.69%
I	7.31%	P	1.82%	V	1.11%
O	7.68%	M	2.61%	X	0.17%
T	9.10%	H	5.92%	Z	0.07%
N	6.95%	G	2.03%	J	0.10%
S	6.28%	B	1.49%	Q	0.11%
L	3.98%	F	2.30%		

Source: *English letter frequency (based on a sample of 40,000 words).* Ithaca, New York: Department of Mathematics, Cornell University. Retrieved September 2015 from http://www.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html

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.

Diphthongs and digraphs to be included will vary by language and by curriculum expectations of each grade level in the country. Using the frequency analysis of sampled textbook pages, the developers identify the most common diphthongs and digraphs students are expected to be able to read. Language experts are consulted to decide the most appropriate letters to include from the curriculum and language frequency analysis.

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 whether or not a child can state a letter sound, allowing for 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):

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

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 is necessary to measure fluency.

Item construction. This subtask consists of 100 total items. Letters of the alphabet, plus any digraphs and diphthongs if appropriate, 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).[14] In other words, each item in the grid remains in the same row in which it appeared in the previous instrument.

Exhibit 12 is one sample of a design for the letter sound identification version of this subtask; **Exhibit 13** shows a mixture of letters and digraphs/diphthongs that was selected for an EGRA in the Wolof language as used in Senegal.

[14] 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: Letter sound identification (Icibemba language, Zambia)

Sub-test 2. LETTER SOUND IDENTIFICATION | **Page 1** | 60 seconds

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. 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.

[point to the letter T] **Icilangililo, iciunda ca cilembo ici t, ni /t/** For example, the sound of this letter is /t/.

[point to the letter M] **Natweshe ukucita ifi: Njebako iciunda ca cilembo ici:** Let's practice: Tell me the sound of this letter.

✓ **Eya cawama, iciunda ca cilembo ici ni /m/.** Good, the sound of this letter is /m/.

✗ **Iciunda ca cilembo ici ni /m/.** The sound of this letter is /m/.

[point to the letter S] **Nomba natwesha icilembo cimbi: Njebako iciunda ca cilembo ici.** Now let us try another one. Tell me the sound of this letter.

✓ **Eya cisuma, iciunda ca cilembo ici ni /s/.** Good, the sound of this letter is /s/.

✗ **Iciunda ca icilembo ici ni /s/.** The sound of this letter is /s/.

[point to first letter] **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.** 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.

Start the timer when the child reads the first letter.

If a child hesitates or stops on a letter for 3 SECONDS, point to the next letter and say "Go on"

When the timer reaches 0, say "stop."

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.

(/) 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

Examples: t m s

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	ŋ	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	ŋ	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)*

Time remaining on stopwatch at completion (number of SECONDS) ☐

Exercise discontinued because the child had no correct answers in the first line ☐

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

Exhibit 13. Sample of letter sound identification subtask with digraphs/diphthongs (Wolof language, Senegal)

Misaal : **o** **uu** **t** **mb**

a	n	o	m	i	u	d	e	L	k
g	b	y	a	uu	x	t	w	n	f
s	r	c	j	p	ñ	oo	i	à	m
n	k	aa	r	ã	ee	ée	d	u	r
ee	ii	b	L	g	y	oo	L	a	t
óo	k	u	a	m	x	é	t	n	c
i	g	ë	j	d	a	e	aa	ng	o
y	x	a	L	m	nd	t	s	nj	L
ŋ	w	u	b	ë	n	i	a	y	e
i	a	aa	k	d	U	o	mb	e	i

Misaal **di** **bu** **dem**

Second Approach: Letter Name Identification

This subtask is very similar in structure and administration to the letter sound identification subtask described above. Although letter sound knowledge is a prerequisite to decoding, numerous studies conducted in the US and European countries have found letter name knowledge to also be highly predictive of later reading achievement. In many countries, EGRA pilot tests of this subtask assessing children's knowledge of letter names have resulted in significant ceiling effects (i.e., almost all children receive high scores), so other subtasks are often used in the final version of the instrument.

Data. As with the letter sound identification exercise, the child's score for this subtask is calculated based on the number of correct letters named per minute.

Item construction. The letter-name version of the subtask is constructed based on the same letter-frequency analysis as described above under the letter-sound version, except this version does not include digraphs and diphthongs. The student stimulus sheet is a laminated page containing all the letters of the alphabet, distributed randomly, 10 to a line in 10 horizontal rows (100 letters total), and evenly distributed among upper- and lowercase letters. Most of the letters will be presented multiple times, based on the frequency with which letters are used in the language. The items within rows of the grid can be reordered (re-randomized) for preparing equivalent test forms, although testing for ordering effects is recommended.

See the letter-sounds sample in Exhibit 12 above for an indication of the layout for the letter-name version of the letter identification subtask.

6.2.3 Nonword Reading

Nonword reading is a measure of decoding ability (i.e., the sublexical route of word processing, as presented in Section 4.3.1) 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 nonwords read, total incorrect responses, 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

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.

[point to the word "opa"] **Icilangililo: Ili shiwi lyapangwa ilyakuti: "opa".** For example, this made-up word is: "opa".

[point to the word "toti"] **Natweshe nomba: belenga ili shiwi.** Let's practice: Please read this word.

✓ **Eya cawama, ilishiwi ni "toti".** Good, This made-up word is "toti."

✗ **Ilishiwi lyakupangafye "toti" talipilibula nangu cimo.** This made-up word is "toti."

[point to the word "maba"] **Nomba esha nalimbi: Belenga nalimbi ishiwi ili.** Now let us try another one. Please read this word.

✓ **Ciisuma, ilishiwi lyaku pangafye ni "maba".** Good, This made-up word is "maba."

✗ **Ili shiwi lyaku pangafye ni "maba".** This made-up word is "maba."

[point to first word] **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.** 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.

(/) 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

Examples: opa toti maba

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)

Time remaining on stopwatch at completion (number of SECONDS) ☐

Exercise discontinued because the child had no correct answers in the first line ☐

Start the timer when the child reads the first word.

If a child hesitates or stops on a letter for 3 SECONDS, point to the next word and say "Go on"

When the timer reaches 0, say "stop."

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.

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

6.2.4 Oral Reading Fluency 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 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 (e.g., inflected forms, derivations) 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 in country X), 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, including the reading comprehension component.

Exhibit 15. Sample: Oral reading fluency with comprehension (English)

EGRA (English) Assessment (LANGUAGE)

Sub-test 4a. ORAL READING PASSAGE	⏱60 seconds	Sub-test 4b: READING COMPREHENSION
Page 4 Show the child the sheet in the student stimulus booklet as you read the instructions.	If a child hesitates or stops on a letter for 3 SECONDS, say "Go on"	**XAfter the child is finished reading, REMOVE the passage from in front of the child.** 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.
Abasem tiawa bi ni. Mepɛ sɛ wokenkan no dennennen, ne ntemntem ma me.Wokenkan wie a, mebisabisa nsem bi afa nea woakenkan no ho.Meka sɛ "Fi Ase" a kenkan abasem no senea wubetumi biara.Wudu asemfua bi so na wunnim a, kɔ asemfua foforo so.Fa wo nsateaa si asemfua a edi kan no so.Metumi afi ase? Hyɛ ase. 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.	If a child does not provide a single correct word before the word in a box, say "Thank you!", discontinue this subtask and check the box at the bottom. Do not ask any comprehension questions.	**Afei mebisabisa wo nsem kakra afa abasem a wokenkan no ho.Bɔ mmɔden sɛ wubeyiyi nsemmisa no ano senea wubetumi.Wubetumi de kasa biara a wopɛ ayiyi nsemmisa no ano.** 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.
(/) 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	If a child says "I don't know," mark as incorrect.	(✓) 1 = Correct (✓) 0 = Incorrect (✓) . = No response.

		Questions [Answers]			
There is no doctor in the village where Ama lives. Father is **sick**.	13	**Who is sick?** [Father]	1	0	.
Ama says that when she grows up she will be a **doctor**.	25	**What does Ama want to be when she grows up?** [a doctor]	1	0	.
She will help people who are sick like **father**.	35	**Why does Ama want to be a doctor?** [to help people / to help people who are sick]	1	0	.
Kojo wants to be a teacher. He will teach boys and girls to be **healthy**.	46	**What will Kojo teach boys and girls?** [to be healthy]	1	0	.
Father smiles. He is happy with both of his **children**.	56	**Why is father happy with his children?** [they want to be doctors / they want to help people]	1	0	.

Time remaining on stopwatch at completion (number of SECONDS)	
Exercise discontinued: the child had no correct answers in the first line	

Mo! Woayɛ ade. Yɛnkɔ ɔfa a edi so no so. Good effort! Let's go on to the next section.

6.2.5 Phonological Awareness – Identification of Initial or Final Sounds; Letter Sound Discrimination

As described in Section 4, "Conceptual Framework and Research Foundations," in order to read, each of us must turn the letters we see into sounds, sounds into words, and words into meaning. Successfully managing this process requires the ability to work in reverse; that is, in order to understand the process of moving from letters to sounds to words, students should also grasp that words are composed of individual sounds and understand the process of separating (and manipulating) words into sounds.

As Stanovich (2000) and others have indicated, "children who begin school with little phonological awareness have trouble acquiring alphabetic coding skill and thus have difficulty recognizing words."

As Stanovich (2000) and others have indicated, "children who begin school with little phonological awareness have trouble acquiring alphabetic coding skill and thus have difficulty recognizing words" (p. 393). Research has found that phonological awareness plays an important role in reading acquisition. Testing for and remediating phonological awareness deficits is thus important for later reading development.

EGRA most commonly measures *phonemic* awareness (one aspect of phonological awareness) through the identification or discrimination of initial or final sounds. These approaches are common in tests of early reading, including:

- Dynamic Indicators of Basic Early Literacy Skills (DIBELS) https://dibels.uoregon.edu/ and Dynamic Measurement Group https://dibels.org/
- Test of Phonological Awareness, Second Edition Plus (TOPA-2+) https://www.linguisystems.com/products/product/display?itemid=10293
- Comprehensive Test of Phonological Processing, Second Edition (CTOPP-2) http://www.pearsonclinical.com/language/products/100000737/comprehensive-test-of-phonological-processing-second-edition-ctopp-2-ctopp-2.html#tab-details

First Approach: Initial or Final Sound Identification

The first approach to assessing phonemic awareness is to have students identify the first (or last) sound in a selection of common words. The example in **Exhibit 16** uses 10 sets of simple words and asks students to identify the initial sound in each of the words. The assessor reads each word aloud twice before asking the student to identify the sound.

Data. The examiner records the number of correct answers. This is not a timed segment of the assessment, nor does it use a pupil stimulus sheet.

Item construction. Simple words are selected from first- or second-grade word lists. If feasible for a given language, only one-syllable words are used, so as not to overtax students' working memory.

Exhibit 16. Sample: Phonemic awareness – Initial sound identification (English)

SUBTASK 2. PHONEMIC AWARENESS | X | Untimed

This is a listening exercise. I want you to tell me the first sound of each word. For example, in the word "pot", the first sound is /p/. I would like you to tell me the first sound you hear in each word. I will say each word two times. Listen to the word, then tell me the very first sound in that word.

Let's practice. What is the first sound in "mouse"? ... "mouse"?

[If the child responds correctly, say:] **Very good, the first sound in "mouse" is /m/.**
[If the child does not respond correctly, say]: **Listen again: "mouse". The first sound in "mouse" is /mmm/.**

Now let's try another one: What is the first sound in "day"? ... "day"?

[If the child responds correctly, say]: **Very good, the first sound in "day" is /d/.**
[If the child does not respond correctly, say]: **Listen again: "day". The first sound in "day" is /d/.**

Ready? Let's begin.

	Item	Answer	Correct	Incorrect	No response
1.	What is the first sound in "at"? ..."at"?	/a/			
2.	What is the first sound in "so"? ..."so"?	/s/			
3.	What is the first sound in "chalk"? ..."chalk"?	/ch/			
4.	What is the first sound in "very"? ..."very"?	/v/			
5.	What is the first sound in "car"? ..."car"?	/k/			
6.	What is the first sound in "for"? ..."for"?	/f/			
7.	What is the first sound in "man"? ..."man"?	/m/			
8.	What is the first sound in "ox"? ..."ox"?	/o/			
9.	What is the first sound in "yes"? ..."yes?	/y/			
10.	What is the first sound in "go"? ..."go"?	/g/			

Exercise discontinued because the child had no correct answers in the first line

Read the instructions to the child and conduct the examples.

Read the prompt and then pronounce the word a second time. Pronounce each word slowly.

If the child does not respond after 3 seconds, mark as "No response" and say the next prompt.

If the child responds incorrectly or does not respond to the first five words, say "Thank you!, discontinue this subtask, check the box at the bottom of the page, and go on to the next subtask.

Second Approach: Letter Sound Discrimination

The second approach involves asking students to listen to a series of three words and to identify which word starts (or ends) with a sound that is different from the others in the series.

A typical design involves 10 sets of three words each. In each set, two words begin with the same sound, and the initial sound for the third word is different. The position of the "different" word in the set varies. The assessor reads each group of three words aloud slowly, two times, and asks the child to choose the word that begins with a different sound. Because this type of task may be completely unfamiliar to children, the protocol for the subtask includes practice sets for the child to try before the assessor begins the actual test.

Data. The examiner records the number of correct answers. This is not a timed segment of the assessment.

Item construction. Simple words are selected from first- or second-grade word lists. As with the initial letter sound approach, words with one or few syllables are used, so as not to overtax students' working memory.

Exhibit 17 is a sample of the letter sound discrimination subtask from an EGRA in Bahasa Indonesia.

Exhibit 17. Sample: Phonemic awareness – Letter sound discrimination (Bahasa Indonesia)

Bagian 2: Membedakan Bunyi Awal

Kegiatan ini tidak dihitung waktunya dan **tidak ada lembar jawab siswa**. Bacakan pertanyaannya terlebih dahulu, lalu bacakan masing-masing kelompok kata yang diteskan sebanyak dua kali. :

Sekarang kegiatan mendengarkan. Saya akan menyebutkan tiga kata. Satu dari ketiga kata tersebut diawali dengan bunyi yang berbeda. Sebutkanlah kata mana yang diawali dengan bunyi yang berbeda itu. Saya akan menyebutkan kata-katanya sebanyak dua kali.

Contoh: **Kata manakah yang diawali dengan bunyi yang berbeda?**
"main" "minum" "kursi".
"main" "minum" "kursi".
Jika benar: **Bagus, "kursi" diawali dengan bunyi yang berbeda.**
Jika salah: **"Main" "minum" "kursi". "Kursi" diawali dengan bunyi yang beda dari "main" dan "minum".**

Kata manakah yang diawali dengan bunyi yang berbeda?
"duku" "jambu" "jeruk".
"duku" "jambu" "jeruk".
Jika benar: **Bagus, "duku" diawali dengan bunyi yang berbeda.**
Jika salah: **"Duku" "jambu" "jeruk". "Duku" diawali dengan bunyi yang beda dari "jambu" dan "jeruk".**

Sekarang coba satu lagi: Kata manakah yang diawali dengan bunyi yang berbeda?
"nenek," "rumah" "nasi".
"nenek," "rumah" "nasi".
Jika benar: **Bagus, "rumah" diawali dengan bunyi yang berbeda.**
Jika salah: **"Nenek" "rumah" "nasi". "Rumah" diawali dengan bunyi yang beda dari "nenek" dan "nasi".**

Apakah kamu siap? Mari kita mulai!

- Ucapkan masing-masing kelompok kata **dua kali** dengan kecepatan sedang (1 kata per detik). Jika anak tidak menjawab dalam 3 detik, tandai dengan tidak menjawab dan lanjutkan.

Berhenti: Jika anak menjawab lima kelompok kata pertama dengan **salah** atau **tidak menjawab,** hentikan kegiatan ini, tandai kotak di bagian bawah , dan lanjutkan pada kegiatan berikutnya.

Katakan, **Kata manakah yang dimulai dengan bunyi yang berbeda?** (Untuk memulai setiap nomor!)

				Jawaban	Tanggapan		
1	sangat	semua	dapur	*dapur*	benar	salah	tidak menjawab
2	kulit	galak	kaget	*galak*	benar	salah	tidak menjawab
3	buku	padi	bias	*padi*	benar	salah	tidak menjawab
4	libur	riuh	ramai	*libur*	benar	salah	tidak menjawab
5	kalah	malas	marah	*kalah*	benar	salah	tidak menjawab
6	guru	mangga	gigi	*mangga*	benar	salah	tidak menjawab
7	jawab	jumlah	cerdas	*cerdas*	benar	salah	tidak menjawab
8	bulan	pesan	pulang	*bulan*	benar	salah	tidak menjawab
9	tidak	tabu	sibuk	*sibuk*	benar	salah	tidak menjawab
10	lapar	lemas	hari	*hari*	benar	salah	tidak menjawab

Tandai kotak ini jika kegiatan tidak dilanjutkan karena siswa tidak dapat memberikan jawaban yang benar pada lima kelompok kata pertama: ☐

Terima kasih, mari kita lanjutkan ke bagian berikutnya.

Bahasa Indonesia 4

6.2.6 Familiar Word Reading

Children's reading skills are often assessed using reading lists of unrelated words. This allows for a purer measure of word recognition and decoding skills than does reading connected text, as children are unable to guess the next word from the context when reading lists of unrelated words. For this assessment, familiar words are high-frequency words selected from first-, second-, and third-grade reading materials and storybooks in the language and context.

Data. Similar to the letter sound identification exercise and for the other timed exercises, three variables are collected for calculating this result: total words read, total incorrect words, and time remaining.

Item construction. Word lists for this task are created from national reading textbooks from the grade levels that will be included in the study. Word-frequency analyses of these texts inform the selection of 50 common, familiar, and simple words representing different parts of speech (e.g. nouns, verbs, adjectives, if applicable). To the extent feasible, the pronunciation of the words is unambiguous and familiar in the relevant language or dialect. Words are arranged horizontally with appropriate spacing and clear, familiar (lowercase) font in 10 rows, five words per line. Items appear in random order (not in order of difficulty, length, alphabetic order, etc.) in the grid. The items within rows of the grid can be reordered (re-randomized) for preparing equivalent test forms, although testing for ordering effects is recommended.

Depending on language characteristics, the selected words include a balance between decodable familiar words (e.g., "cat") and common sight words (e.g., "the"), as well as parts of speech. Word length and spelling patterns are representative of those found in early grade readers, and words are composed of a variety of letters, with none repeated disproportionally. None of the items can be a word in any other language with which the children may be familiar. This task must not include one-letter words, as these will already be included in the letter grid. Three additional words serve as example words for assessors to practice with students. The words must be similar in level of difficulty to the words in the grid.

The font used in this subtask is similar in size and style to that used in the official reading textbooks or, if there is no official book, in the most common books available for purchase. See also the brief discussion about fonts in the letter sound identification section (Section 6.2.2).

Exhibit 18 is a sample of a familiar word reading subtask from an EGRA administered in Portuguese in Timor-Leste.

Exhibit 18. Sample: Familiar word reading (Portugese, Timor-Leste)

Sesaun 4. Identifika Liafuan ne'ebé Familiar

Hatudu ba labarik papél ida ho liafuan familiar sira iha pájina daruak (formuláriu ba labarik). Dehan:

Iha ne'e iha liafuan balun. Ha'u hakarak ó atu lee mai ha'u liafuan sira ne'ebé mak ó bele lee (labele soletra maibé lee de'it). Ezemplu, liafuan ida ne'e: "gato".

Mai ita koko: favór lee liafuan ne'e [hatudu ba liafuan "encarnado"]:

Sé labarik hatán ho loos dehan: **Di'ak, liafuan ida ne'e "encarnado"**

Sé labarik hatán laloos dehan: **Liafuan ida ne'e "encarnado"**

Agora ita koko fali seluk: favór lee liafuan ne'e [hatudu ba liafuan "cantar"]:

Sé labarik hatán ho loos dehan: **Di'ak, liafuan ida ne'e "cantar"**

Sé labarik hatán laloos dehan: **Liafuan ida ne'e "cantar"**

Bainhira ha'u dehan "Hahú", favór temi sai letra nia naran hotu ne'ebé mak ó hatene. Lee letra sira iha pájina nia naruk, komesa husi risku dahuluk nia okos. *[hatudu ba iha letra dahuluk iha liña dahuluk depois ezemplu].*

Prontu? Komesa.

Hahú ho kronómetru bainhira labarik komesa lee letra dahuluk.

*Haree no dada tuir imi- nia lapis atu marka loloos letra ruma ne'ebé mak laloos ho barra **(/).***

*Kuandu labarik kuriji, konta ida ne'e iha loos. **Nonok nafatin,** ezetu ho situasaun hanesan; bainhira labarik nonok liu segundu/detik 3, dehan sai letra nia naran no marka letra ne'e laloos, hafoin kontinua hatudu fali letra tuir mai no dehan **"favór bá oin."***

Depois segundu 60 ona dehan, "PARA".

Marka letra ne'ebé lee ikus ho kolxete (]).

***Regra atu hapara ezersísiu**: sé labarik la fó resposta ida ka rua ne'ebé loos iha liña dahuluk, dehan "**obrigada barak",** ba ezersísiu ne'e, marka kaixa ida iha okos hafoin pasa fali ba ezersísiu tuir mai.*

Exemplo : gato encarnado cantar

1	2	3	4	5	
ir	embora	azul	amar	vir	(5)
ajudar	dois	correr	ver	abaixo	(10)
encarnado	bola	brincar	acima	vaca	(15)
dever	querer	são	agora	baixo	(20)
favor	cedo	gostar	eles	bom	(25)
obrigado	vindo	quando	saber	ele	(30)
pular	gato	uma	voar	poder	(35)
porque	verde	cantar	aqueles	sempre	(40)
várias	qual	sorriso	sentar	limpar	(45)
sete	beber	casa	eu	junto	(50)

Marka **X** iha kaixa tuir mai ne'e sé ezersísiu la kontinua tanba labarik la hatán loloos iha liña dahuluk: ☐

Sé labarik lee kompletu letra hotu iha ezersísiu ne'e maibé menus husi segundu 60 entaun hakerek segundu hira mak sira uza ba lee: _____

Halo nota ba totál númeru letra sira durante tempu ezersísiu (segundu 60): _____

Halo nota ba letra ne'ebé LOOS durante tempu ezersísiu: _____

Halo nota ba letra ne'ebé LALOOS durante tempu ezersísiu: _____

Servisu di'ak! Mai ita kontinua sesaun tuir mai.

6.3 Review of Additional Instrument Components

As mentioned above, several other less commonly used subtasks have been created and piloted in EGRA instruments, depending on factors such as language idiosyncrasies, specific research questions, or curriculum-related questions. They are described briefly below.

6.3.1 Dictation

Dictation assessment is frequently used by teachers to test both oral comprehension and writing skills. As discussed earlier, the reading process can also be tested in reverse: Students' ability to hear sounds and correctly write the letters and words corresponding to the sounds they hear demonstrates their success with the alphabetic principle. A number of assessment packages offered by commercial test development specialists give teachers instructions on how to develop and score their own assessments. This particular subtask of the EGRA was inspired by models promoted in the early 2000s by the Educational Testing Service and the Children's Literacy Initiative (neither model remains available) and by research by the International Reading Association (Denton, Ciancio, & Fletcher, 2006). This subtask was part of the original core EGRA; however, it was later removed due to difficulties in standardization of scoring and implementation. As of 2015, it was being used in some contexts, but not widely.

Data. Students are scored on a simple scale that captures the students' ability to correctly write letters. The capital or lowercase versions of the letter are both acceptable answers.

Item construction. Five of the most commonly occurring letters of the language being assessed are chosen for this subtask.

Exhibit 19 is a sample dictation subtask in two parts that involved writing individual letters, then writing several short words (as opposed to sentences).

Exhibit 19. Sample: Dictation – letter writing (Creole, Haiti)

K-Seksyon 8 : Dictée	Fèy papye ak kreyon	×
	×	×

(✓) Korèk / pa korèk / pa gen repons ditou

A. Lèt

Bay timoun nan yon kreyon ak yon fèy papye. Pa kite l gade lèt yo. Si timoun nan di : « Mwen pa konnen, » make repons sa a kòm enkòrèk.

Mwen pra l di w kèk lèt. Se pou koute m avèk atansyon. Apre chak lèt mwen fin di w, m ap repete l yon lòt fwa pou ou, e w ap ekri lèt ou tande a sou papye a pou mwen. Eske w konprann sa m mande w fè a ? Oke, koute epi ann kòmanse.

[Li chak let 2 fwa]			
b	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
j	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
m	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
v	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
z	○ Kòrèk	○ Pa Korèk	○ Pa gen repons

Kounye a, mwen pra l di w kèk mo. Koute m avèk atansyon. Apre chak mo mwen fin di w, m ap repete l yon lòt fwa pou ou, e w ap ekri mo ou tande a sou papye a pou mwen. Eske w konprann sa m mande w fè a ? Oke, koute epi ann kòmanse.

B. Mo

[Li chak mo 2 fwa]			
fil	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
ten	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
pay	○ Kòrèk	○ Pa Korèk	○ Pa gen repons

Anfòm! Ou pare pou w fè pwochen aktivite a. Trè byen !

6.3.2 Phoneme Segmentation

Phoneme segmentation, which in this case means the division of words into phonemes, is one of the most complex skills of phonological awareness and should be emphasized in the early grades (Linan-Thompson & Vaughn, 2007). It is also predictive of later learning skills. EGRA subtasks designed to test phoneme segmentation have tended to be difficult to administer and have demonstrated large floor-effect problems (i.e., very few students are able to complete the subtask).

The original EGRA subtask for phonemic awareness involved phoneme segmentation. For this portion of the assessment, the examiner would read aloud a list of 10 simple, one-syllable words, one at a time. Students were asked to identify and sound out each sound present in the word (as this was an auditory assessment, there was no student handout). This task was later removed from the core list due to floor effects. It has been used in some contexts more recently where students have more familiarity with phonemic awareness.

Data. The examiner records the number of correct phonemes as a proportion of total phonemes attempted. This is not a timed segment of the assessment.

Item construction. Simple one- or two-syllable words are selected. Words use common phoneme constructions that minimize the number of complex graphemes (i.e., phoneme constructs with more than one letter and with blends are avoided) and vary the beginning sounds between consonants and vowels.

Exhibit 20 is an example of a phoneme segmentation subtask from a 2007 EGRA in Portuguese in Timor-Leste.

Exhibit 20. Sample: Phoneme segmentation (Portugese, Timor Leste)

Sesaun 3b. Kualidade ba Fonema sira - 2

Ida ne'e "LA'OS" ezersísiu atu marka ho tempu ***NO LA EZISTE FORMULÁRIU BA LABARIK.*** *Lee ho lian maka'as kada liafuan ida-idak dala rua hafoin husu labarik atu dehan ninia son sira.*

Ida ne'e ezersísiu kona-ba rona. Ó hatene katak kada letra iha nia son ida-idak. Ezemplu, "casa", "c"-"a"-"s" -"a" bele rona "/c/ - /a/ - /s/- /a/". Ha'u sei temi liafuan ne'e dala rua, entaun rona didi'ak.

Ita koko. Son saída mak iha liafuan "par" – "par"?
[sé labarik hatán loos, dehan]: **Di'ak, son ba liafuan "par" mak /p/ /a/ /r/.**
[sé labarik hatán laloos, dehan]: **son ba liafuan "par" mak /p/ /a/ /r/.**
Agora ó fali. Dehan mai ha'u son iha liafuan "par". *[hein ba labarik to'o segundu 5 atu hatán].*

Ita koko fali seluk. Son saída mak iha liafuan "mar" – "mar"?
[sé labarik hatán loos, dehan]: **Di'ak, son ba liafuan "mar" mak /p/ /a/ /r/.**
[sé labarik hatán laloos, dehan]: **son ba liafuan "par" mak /p/ /a/ /r/.**
Agora ó fali. Dehan mai ha'u son iha liafuan "par". *[hein ba labarik to'o segundu 5 atu hatán].*

Agora ita komesa. Ha'u sei temi liafuan ida – ida dala rua. Rona ba liafuan ida-ida no dehan mai ha'u liafuan ne'e nia son. Ó hatene ona atu halo saidá?

Pronunsia ***DALA RUA*** *ho neineik kada liafuan ida-idak ninia konjuntu (liafuan ida segundu ida).*

*Marka ho barra (**/**) ba resposta ne'ebé laloos, no moos ba son ne'ebé sira la temi ou hakat liu.*

Sé labarik la hatán ba liafuan ida iha segundu 5 nia laran, entaun marka son sira iha liafuan ne'e laloos no kontinua ba liafuan seluk.

Regra atu hapara ezersísiu: *sé labarik la fó resposta ida ka rua ne'ebé loos iha liafuan 5 dahuluk, dehan "**obrigada barak",** keta kontinua ezersísiu ne'e, marka kaixa mamuk ida iha okos hafoin pasa fali ba ezersísiu tuir mai.*

Son saída mak iha _____? _____/ [*repete liafuan ne'e dala rua*]]					
"lar"	/l/	/a/	/r/		_____/3
"era"	/e/	/r/	/a/		_____/3
"paz"	/p/	/a/	/z/		_____/3
"sal"	/s/	/a/	/l/		_____/3
"mal"	/m/	/a/	/l/		_____/3 ***Linha 5***
"arroz"	/a/	/r/	/o/	/z/	_____/4
"casa"	/c/	/a/	/s/	/a/	_____/4
"fala"	/f/	/a/	/l/	/a/	_____/4
"massa"	/m/	/a/	/s/	/a/	_____/4
"carro"	/c/	/a/	/r/	/o/	_____/4

*Marka **X** iha kaixa tuir mai ne'e sé ezersísiu la kontinua tanba labarik la hatán loloos iha liafuan 5 dahuluk: ☐*

Servisu di'ak! Mai ita kontinua sesaun tuir mai.

6.3.3 Maze and Cloze

Maze and cloze assessments are not uncommon in classroom settings as a way to test comprehension. *Maze* involves creating a brief paragraph of connected text or narrative that tells some sort of story; then replacing every n^{th} word[15] with a "multiple-choice" type option of three possible words, some of which make sense in the context while others may not. The student can read them silently or aloud and selects one word (from the series of three words) that would best complete the phrase. *Cloze* involves the same type of word removal but replaces the missing words with a blank space and allows the student to supply the missing words. This measure has not yet been widely vetted. Cloze can be very difficult to construct in a way that is appropriate for students' reading levels, and it generally is a new or unfamiliar type of task for students, which can cause them to perform poorly. Maze and cloze are not timed; however, they often have a time limit within which students are to complete the subtask (typically 3 to 5 minutes).

Data. Responses are scored based on a key indicating acceptable responses, with a mark for items correct out of the total number of items.

Item construction. The first and last sentence of the paragraph or narrative are complete. The remaining sentences each are missing only one word. However, in a couple of the sentences, a short phrase could be missing. For the sentences, the missing item is not the first word/phrase, and there is only one missing word or phrase per sentence. There are three options for each missing word/phrase. The options in the response set might be plausible, but they are not responses that would actually be possible. So, there really is only one true answer in the response set. For example, in the sentence "The lake was big, and he saw the fish (swimming/receiving/digging)," whereas a loose argument could be made that a fish can "dig" in sand, the more plausible answer in this case would be "swimming."

Exhibit 21 consists of a sample of a maze subtask in English.

[15] It is recommended that the word being replaced not be the first word of the sentence.

Exhibit 21. Sample: Maze (English, Kenya)

Jane does not like to do homework. When she gets home from school (**cat, she, fly**) only wants to play. Jane tells (**table, her, red**) mother that a full day of (**scary, school, house**) and homework is too much! Her (**hen, to, mother**) tells Jane she needs to do (**her, book, run**) homework before playing. Jane tells her (**goat, mother, work**) that she is just a little (**child, leg, three**), so she needs more time to (**eats, hat, play**).

One day, Jane decides she will (**under, never, dog**) do homework again. She does not (**bring, throw, with**) her books home from school anymore. (**She, Class, Jump**) feels that she is on a (**shop, holiday, on**). Jane is happy. After two full (**weeks, chairs, up**), she takes her exams. Jane gets (**dirty, ear, poor**) marks. Her mother is very angry, (**and, in, hot**) Jane is very sad. She cries (**to, for, walk**) a long time.

Her big brother (**on, pulls, comes**) to see her. She tells him (**in, pen, about**) her homework and her very poor (**marks, pigs, fear**). The brother tells her that doing (**shirt, have, homework**) will help her a lot. Now, (**box, run, Jane**) knows that homework can help her (**get, drink, bus**) good marks. She now wants to (**fall, work, pink**) hard and do her homework.

6.4 Reasons for Exclusion of Other Potential Instrument Components

During original instrument development, both the literature review and the expert review process generated numerous suggestions for inclusion of additional test components and measures. As each of these suggestions was reviewed, selection criteria were established for the appropriateness of their inclusion in the instrument. The main consideration was the usefulness of each test in predicting future student success in reading.

Foremost among these suggestions was the inclusion of a picture-based subtask, such as those in the Peabody Picture Vocabulary Test (PPVT), a commercially available test from Pearson Learning Group. Some variants of early grade reading assessment tools (including a version applied by Plan International in French in West

Africa) have included pictures to identify knowledge of common vocabulary (such as body parts: hand, head, toe, etc.).

However, the original EGRA developers omitted pictures or picture vocabulary tests for several reasons: (1) Vocabulary is indirectly measured in both the listening comprehension and paragraph reading segments; (2) development of pictures frequently runs into copyright issues (use of the PPVT, for example, was discarded as an option because copyright permissions would have to be sought each time the instrument was used in another country); and (3) it would have been very difficult to create pictures that would be universally appropriate for all cultures and contexts—or alternatively, to expect to recruit in-country illustrators to create original artwork within the brief time frame allotted for each EGRA adaptation. In addition, when pictures are locally developed and crafted, experience has shown that at least two problems seem to arise. First, the pictures are often of very low graphical quality, making it difficult sometimes for even an experienced adult to interpret the picture and answer the question. Second, even assuming high graphical quality, developing appropriate picture-based items seems to require considerable skill because of the difficulty in having pictures interpreted similarly and acceptably across multiple countries. Issues of cultural relevance make picture-based measures extremely difficult to create and standardize.

Another component initially tested and later eliminated from the assessment was derived from Marie Clay's (1993) Concepts About Print assessment. Early applications of a subtask requiring children to indicate where to begin reading, which direction to read, and where to read next, demonstrated ceiling effects (nearly all children successfully completed the task). Furthermore, deriving conclusions from both United States and international research, the *Handbook of Psychology* reported that print awareness appears to have little predictive power of later reading skills; it mainly serves as a proxy measure for print exposure and literacy environments (Paris & Paris, 2006). Based on these results as well as efficiency and time limitations, the EGRA assessment does not include a "concepts about print" subtask.

6.5 Translation and Other Language Considerations

6.5.1 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.”

6.5.2 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 that 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 across different versions).

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, 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 contrast, 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 alphabetic languages such as Italian, Finnish, and Greek require only about a year of instruction for students to reach a comparable level of reading proficiency (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. In addition to transparency, the rate of acquisition of reading skills is also affected by orthographic complexity. Orthographic complexity includes the number of symbols that need to be learned and their visual complexity. For instance, most South Asian and Southeast Asian orthographies have approximately 500 symbol combinations that need to be acquired (Nag & Perfetti, 2014), and thus the process of acquiring reading ability in these languages may take 4–5 years (Nag, 2007) compared to 2–3 years in deep alphabetic orthographies and 1 year in shallow alphabetic orthographies (Seymour et al., 2003). Visual complexity is also an important factor in orthographic complexity, wherein letters are graphically presented nonlinearly and symbols physically appear below, above, to the left of, or to the right of the sound they sequentially follow. In some cases, such as in Arabic, they do not appear at all visually, and are inferred through context. These factors also impact rate of

acquisition of various reading skills (e.g., Kandhadai & Sproat, 2010; Karanth, 2002), and need to be considered for cross-language comparisons.

Another important factor in cross-language comparisons is differences in phonology. Because some languages represent sounds only as syllables, some only as phonemes, and some as a combination (Perfetti, 2003), it is important to determine which type of phonological awareness measure is most suitable for the phonological structure of the 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[16] 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 6.5.1 above).

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;
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 F.**

[16] A "technically adequate instrument" is one that has been demonstrated to produce reliable results, allows the generation of valid analyses, and therefore lends confidence.

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

In instances in which subtask difficulty from EGRA Instrument A and Instrument B is determined post-test not to be equal, specific test equating procedures should be applied to account for the differences (see Section 10.5). *Equated test forms,* therefore, refers to forms that have been adjusted by a statistical process following test administration 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 6), 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, so that no matter how far a student gets in the assessment before the time is up, his or her experience with that assessment will be the same as with a previous test administration.

While these techniques are intended to limit the need for equating, they do not guarantee equivalent forms, nor do they remove the need to test for equivalence after piloting. For situations in which these techniques are used but still result in non-equivalent test forms, statistical equating methods may be required. Section 10.5 discusses specific methodologies and recommendations for equating scores after data are processed and analyzed.

6.7 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 6 as well as Section 9 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.

7 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 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.

Electronic data collection 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;[17] and the software system Tangerine®, created by RTI

[17] Under a subcontract to RTI International on the USAID EdData II project (see Collins & Messaoud-Galusi, 2012; Prodigy Systems, 2011).

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.[18] 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 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 Early Grade Mathematics Assessment (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.

7.1 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.

[18] Laptop computers were not considered a viable technology for this purpose because of potential mode effects due to the visibility of the technology in the classrooms; and because of limitations on their use in certain contexts (lack of electrical supply, dust/humidity, transportation by foot, bicycle, boat, etc.). Data collection and data entry systems also exist for desktop or laptop computers; for example, eEGRA developed and used by Education Development Center, Inc. (EDC) (http://eegra.edc.org/).

- **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 of 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 opting for 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 global positioning system (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.

7.2 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). Features of Tangerine and several other e-data collection tools—Magpi, SurveyToGo, doForms, Droid Survey, Open Data Kit (ODK), and Command Mobile—are compared in **Annex G,** and a sample of paper versus electronic instructions is presented in **Annex H.** As of this writing, iProSurveyor (for the iPad), Tangerine, and SurveyToGo were the only platforms not including laptop or desktop data entry systems (see footnote 18 on the previous page) 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.

7.3 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.

It is essential, as part of the implementation process, to establish clear procedures for ownership, access, and use of the hardware, software, and data. It is common (and is cost effective) for hardware to be reused by the implementer or funding organization, or for ownership of purchased items to be transferred to local organizations for continued use.

7.4 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

8 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).

The assessors who will be piloting the instrument will need a training of about five working days.[19] 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.

For an EGRA training in Tanzania in 2013 that had 150 assessors, an instrument in two languages, and additional surveys, the technical team of trainers included five people: a specialist in language 1; a specialist in language 2; an expert in the data collection software; a logistics lead; and an overall coordinator who also focused on the assessor performance tests, pre-workshop preparation, survey design, and donor relations.

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.

[19] See Section 9.1.3 on the pros and cons regarding the various possible timings of the assessor training in relation to the pilot and full data collection.

8.1 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 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.

Another factor to be considered at the recruitment stage is whether candidates may be subject to conflicts of interest—in either the public or the private sector—stemming from the current political landscape within the country.

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.

8.2 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 I** 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.

8.3 Components of Assessor Training

As indicated via the sample agenda in Annex I, 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 8.7); use the results for remediation and ultimately for selecting the assessor corps for the main data collection.

8.4 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.

8.5 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.
- 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 ten[20] 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.

[20] 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 to each student (i.e. multiple languages or multiple subjects).

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

A quiet and separate space at the school will be needed for participants to practice administering the assessments. 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.

8.6 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 wise 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
- 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 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.

8.7 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.

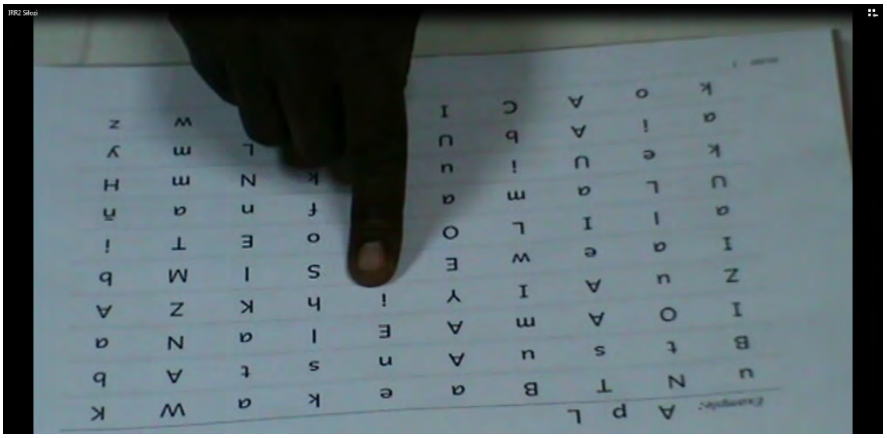
This type of evaluation is particularly helpful for improving the assessors' performance before they get to the field. It must also be used for selecting the best-performing assessors for the final assessor corps for the full data collection, as well as alternates and supervisors.

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 22**), 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 22. 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 J** 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 J for more information about how to evaluate the assessor accuracy data.

In addition to the assessor evaluation process during training, it is required that assessors continue to test the reliability and consistency among themselves (interrater reliability, or IRR) once they are in the field collecting data. IRR can help further the reliability and the consistency of the data as they are being collected, as well as prevent *assessor drift* (see glossary). Additional information on interrater reliability measures during the data collection process is presented in **Exhibit 23; Annex K** contains charts showing several sample plans for varying the assessor pairings.

Exhibit 23. Sample protocol for monitoring interrater reliability during fieldwork

Save the Children®

Protocol for Collecting Interrater Reliability Data

An important part of any data collection process when performing one-on-one assessments is to see how well assessors agree with one another, and how reliably they score students. In an ideal world, assessors would mark every response exactly the same. However, it can happen that assessors disagree about whether to mark a student correct or incorrect. Hopefully, the piloting and training process will help raters to consistently agree with each other. Nonetheless, it is important to continuously measure the rate of agreement between assessors. This is done according to the following procedure.

Each day, assessors work as a team to assess of the first student of the day. For example, if a team of assessors has 4 individuals, then Assessor A and Assessor B are a team. Students are randomly sampled as normal. Assessor A calls the first student and brings the student to the office/tree/location of the assessment, where Assessor B is also waiting. Assessor B sits in a position from which he/she cannot see what Assessor A writes. Assessor A conducts the assessment as normal, asking the background questions and the various reading and/or numeracy skills, while recording the student's responses. Assessor B begins scoring a separate assessment for the same student. During the assessment, Assessor B never asks any questions,

but merely listens and records. Hence, ***two* assessments are recorded for the first student of every school**. Assessors C and D will follow this same procedure with another student, recording two more assessments for the second student assessed in the school.

Assessors must be careful to indicate on the assessment whether they are administering the assessment or whether they are only listening and recording. This requires an item on each assessment where assessors can record this information.

Once the assessment of the first student is finished, Assessor A thanks the student for participating and sends that student back to class. Then Assessor A and Assessor B compare how they scored the student. Assessors A and B should discuss any items on which they disagreed, and resolve the proper way to have scored that particular item. If the 2-person team cannot resolve their scoring disagreement, it should be noted, and brought to the attention of the entire group of assessors at the end of the day. Please note: **Once Assessor A and B enter a response onto their assessment, it should never be changed, erased, or corrected after the student has left the room.** These points of disagreement are important to retain, as they will provide information on inter-assessor agreement and reliability. It is perfectly natural for there to be some disagreement between assessors. Measuring the amount of disagreement is important in the data analysis process, as it will provide information on how much assessor measurement error might affect the observed reading scores of students.

Once the assessors have discussed their matching assessments, they should separate and each call their next student for individual assessment.

At the subsequent schools, the teams should be altered, so that different members take on the role of talking/listening assessors, and that each assessor is paired with a different assessor each day.

© 2015 by Save the Children. Used by permission. All rights reserved.

9 FIELD DATA COLLECTION: PILOT TEST AND FULL STUDY

9.1 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.[21]
- 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.

[21] 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 6.6 for guidelines on creating equivalent instruments and Section 10.5 for guidelines on statistical equating.

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.

Exhibit 24 outlines the key differences between the pilot test and the full data collection.

Exhibit 24. Differences between EGRA pilot test and full data collection

	Pilot test	Full data collection
Purpose:	To test the reliability, validity, and readiness of instrument(s) and give assessors additional practice	To complete full assessment of sampled schools and pupils
Timing:	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
Sample:	Convenience sample based on target population for full data collection	Based on target population (grade, language, region, etc.)
Data:	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
Instrument revisions:	Can be made based on data analysis, with limited re-piloting after the changes	No revisions are made to the instrument during data collection

9.1.1 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 nonzero scores must be of a reasonable range and comparable to the nonzero 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 threefold. 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 nonzero scores. For example, if it is anticipated that only 20% of cases would result in nonzero scores, a pilot sample of 750 students would be required to obtain the 150 nonzero scores needed for psychometric analysis. However, oversampling of 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 socioeconomic information for the specific geographic areas and use this information as a proxy for school performance levels. It is not recommended, however, that the convenience sample include 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 nonzero scores may be quite different.

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.

9.1.2 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.[22]

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. Some other types of reliability tests are described in Annex E.

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. 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 exploratory factor analysis. 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

[22] 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.

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 (i.e., 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 8, specifically Exhibit 23, "Sample protocol for monitoring interrater reliability during fieldwork." 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 ongoing 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 J, Section J.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 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 (≥ 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.

9.1.3 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.

9.2 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 responsibility (quality, confidentiality, and anonymity[23]).
- Monitor IRR data to improve the quality of data and reduce the potential of "drifting"— (also known as *assessor drift*—see glossary)
- 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
- A laminated book of student stimuli, one per assessor (the same laminated book will be used for each student that the assessor tests)[24]
- Stopwatches or timers (in case tablets fail and backup paper instruments must be used)
- Pencils with erasers, and clipboards

[23] Anonimity: 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.

[24] 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).

- 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 probing 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 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 8.7 and Annex K.

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 (assessor 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 able to hear or see the testing.

9.3 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.

9.3.1 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.

9.3.2 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 25** on the following page presents the considerations required.

Exhibit 25. Determinants of the sampling groups			
Research design—teacher data:	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
Basis for sampling—grade or class:	Grade level	Class level – one class per grade	Class level – more than one class per grade
Notes: • Surveys may involve one or more grades. • In addition to selection by grade/class, the research design may specify that the students be selected by sex (see next row). • 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.			
Group(s) from which the sample(s) must be selected:	Either: • All the students in each grade (irrespective of gender) Or: • All the male students in each grade, and • All the female students in each grade	Either: • All the students from each selected class in each grade (irrespective of gender) Or: • All the male students in each selected class, and • All the female students in each selected class	

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.[25]

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 four groups and five students (20 ÷ 4) 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*

[25] See Annexes B and C, and Section 5, for more information on sample design.

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.
    - 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 pth 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 ÷ 8 = 6.75; round: p = 7*
3. *Randomly select a child from the first p = 7 children*[26] *– for example, child number 3*
4. *Select every* p^{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 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).

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.

9.4 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

[26] This process is known as "random start."

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.

9.5 Uploading Data Collected in the Field

Assuming data are collected electronically (this is current recommended best practice—see Section 7), 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.

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 see uploaded data, and whether 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.

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.

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.

10 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 and assessments, the sample methodology, and the data structure, and then 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.

10.1 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 26 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 26. 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 assessments could be verified as accurate and as 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 10.2.

☐ **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.

10.2 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 L** 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 27** provides a list of variable names for EGRA subtasks as well as the variable names for timed scores (if the subtask is timed).

Exhibit 27. EGRA subtask variable nomenclature and names of the timed score variables

Name of subtask variable	Name of subtask	Name of subtask timed variable	Name for subtask timed
letter	Letter Identification (Names)	clpm	Correct Letter Names per Minute
letter_sound	Letter Identification (Sounds)	clspm	Correct Letter Sounds per Minute
fam_word	Familiar Word Reading	cwpm	Correct Words per Minute
invent_word	Nonword Reading	cnonwpm	Correct Nonwords per Minute
oral_read	Oral Reading Fluency	orf	Oral Reading Fluency
read_comp	Reading Comprehension		

Exhibit 27. EGRA subtask variable nomenclature and names of the timed score variables

Name of subtask variable	Name of subtask	Name of subtask timed variable	Name for subtask timed
list_comp	Listening Comprehension		
syll_sound	Syllable Identification (Sounds)	csspm	Correct Syllable Sounds per Minute
oral_vocab	Oral Vocabulary		
vocab	Vocabulary		
maze	Maze		
dict	Dictation		

10.2.1 <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.

10.2.2 <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, for example, 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 in 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 28 contains some examples of how the EGRA variables are named, based on the language and the number of sections repeated within the instrument.

Exhibit 28. 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_pcnt	Percent Correct	0-100
_score_zero	Zero Score Indicator	0 "Score>0" 1 "Score=0"
_attempted_pcnt	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_pcnt.** 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_pcnt.** _score divided by _attempted.

10.3 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 is 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 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 a 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:

$$\text{Subtask_per_minute} = \frac{\text{Subtask_score}}{\text{Time given for subtask-subtask_time_remain}} \times 60$$

10.4 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_pcnt80,** 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).

10.5 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 children being assessed with one form would have received had they been assessed with a different test form (Holland & Dorans, 2006; Kolen & Brennan, 2004).

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 techniques that can be used for statistical equating 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) ideally should account for at least 20% to 25% of the total items on the assessment and represent a mini-version of the overall assessment (in terms of difficulty and variation).[27] 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., nonanchor 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).

[27] There is some debate about the exact proportion of required anchor items, but 20% to 25% is an oft-cited guideline.

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.[28]

In this scenario, it is suggested that EGRA developers create 10 reading comprehension passages with five questions on each (10 sets), using expert judgement 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 29** (with 10 forms of 3 sets and 15 questions, each).

Exhibit 29. Sample counterbalanced design

Number of students	First set	Second set	Third set	Pilot test forms, by letter
50	1	2	4	A
50	2	3	5	B
50	3	4	6	C
50	4	5	7	D
50	5	6	8	E
50	6	7	9	F
50	7	8	10	G
50	8	9	1	H
50	9	10	2	I
50	10	1	3	J
500				

In this design, every passage appears in each set (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 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 equating: Equating models based on CTT establish relationships between total scores on different test forms. This is a more "traditional"

[28] 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 costs versus benefits, and the value of having 10 evaluated passages.

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 equipercentile equating. This toolkit does not provide in-depth explanations of each approach but there are additional recommendations regarding these approaches in **Annex M.**

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. In the context of the EGRA subtasks, this approach is most useful for equating oral reading fluency.

Item response theory 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 for 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 (via Rasch models) can be particularly beneficial for pre-equating on subtasks that have few items (a shortcoming of the CTT equating approaches), as long as those subtasks have useful item-level data. Such subtasks include reading comprehension, listening comprehension, dictation, vocabulary, and maze.

Areas for further deliberation regarding test equating procedures are included in Annex M.

10.6 Making EGRA Data Publicly Accessible

USAID is expected to make public-use files (PUFs) containing early grade reading assessment data publicly available through the Secondary Analysis for Results Tracking Education portal (SART Ed) and the Development Data Library (DDL), and to an increasing degree, via the World Bank's EdStats platform. Other funders may choose to follow a similar process for data collection that they sponsor. Public-use

files are cleaned, finalized, and de-identified data sets intended for public consumption. These data sets contain all relevant variables needed for proper data analysis, but have all identifiable information masked to protect the identities of individuals and establishments. USAID's Update to Reporting Guidance defines cleaned, finalized, and de-identified data as follows:

- **Cleaned data.** Implementer checked for and corrected any apparent errors or inconsistencies in content, missing information, etc.
- **Finalized data.** Data sets include any derived or secondary indicators that the implementer used to calculate indicator values that were included in reports. The implementer finished processing the data set and no further changes are anticipated.
- **De-identified data.** Steps have been taken to protect the privacy and anonymity of individuals and schools associated with an assessment. The implementing organization worked with its Institutional Review Board to ensure that assessment participants are properly protected.

Annex N includes further recommendations and guidelines on how to clean, finalize, and de-identify data such that it can be distributed widely to public audiences. Once a PUF of a data set has been created, well-documented information is needed to help the public users familiarize themselves with the data. For EGRA surveys conducted with USAID funding, the following information is provided to the users:

- Background information, such as definition of the population of interest—including the source of the sampling frame used to draw the sample; description of the sample design; and time of data collection.
- All relevant documentation, including the questionnaires and assessment tools used. (If the EGRA was conducted for program impact evaluation, the questionnaires are released only after the program is complete, so as not to compromise materials that may be used for future EGRA studies.)
- The written data analysis report submitted to USAID and approved.

Implementers should recognize the importance of documenting the names and descriptions of key variables along with the settings needed for proper data analysis. Specific guidelines on how to include variable names and descriptors for purposes of data analysis are included in Annex N.

At the time this edition of the toolkit was being written, several of USAID's data repositories were in the development stage. However, it is still important for early grade reading assessment data to be publicly available. Implementers who have collected EGRA data are therefore recommended to:

- Post the PUF in an accessible location online, accompanied by easily locatable documentation (e.g., all items are located in one zipped file or the website contains links to these documents).
- Create the PUF using a nonproprietary data file and, when possible, a proprietary data file ready to be analyzed (that is, with the complex survey design already specified).
- For a nonproprietary file, create a csv, comma-separated values text file.
- For a proprietary file, generate either a Stata .dta file or an SPSS .sav file (along with the SPSS .csaplan file).

11 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.

11.1 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.

11.2 Inferential Statistics

Inferential statistics allow evaluators to reach conclusions about entire populations based on a sample of that population, to draw inferences about hypotheses regarding population parameters, and to compare two different populations (e.g., treatment and control groups). Inferential statistics are critical for both impact evaluations and snapshot EGRAs that seek to make statements about education for an entire country or region based on a sample of students or schools in that country or region. The type of inferential statistics needed for a given student depends on the evaluation design, as follows.

- **Experimental designs (or randomized controlled trials).** Two groups can be compared using paired *t*-statistics to determine, for example, whether endline scores were higher for treatment group participants than for control group participants. If randomization was successful and the sample size is large enough, it is not necessary to also consider differences between the treatment and control groups in terms of demographic factors or baseline scores, as both groups would be identical due to the randomization process.
- **Quasi-experimental designs using a longitudinal design (tracking individual students over time) or semi-longitudinal design (tracking teachers or schools over time).** If all the following conditions were met, then evaluators can also use *t*-statistics and/or gain scores to show changes over time: (1) Treatment and comparison groups were well matched at baseline, (2) there were no significant differences between the two groups, and (3) schools or students were tracked over time in both groups through a longitudinal design. Otherwise, quasi-experimental analyses (such as those described below) are necessary to accurately measure growth and/or program effectiveness.
- **QEDs using a cross-sectional design.** To compare two groups, evaluators must use a quasi-experimental analysis approach such as difference-in-differences (DID), regression discontinuity, instrumental variables, or regression analysis (preferably with a balancing variable). DID subtracts baseline outcomes from endline outcomes (creating gain scores) for both the treatment and comparison groups and then subtracts the comparison group gain score from the treatment group gain score to get the treatment effect. This approach is intended to account for baseline differences but still relies on the assumption that performance trajectories before baseline were consistent across groups. It is often beneficial to pair this approach with a matching procedure (such as

propensity score matching) in order to ensure that the two groups are as similar as possible. When using regression analysis, evaluators must include time and treatment dummies as well as an interaction dummy for time × treatment to determine the effect (which is essentially the regression approach for a DID estimate). Additional independent variables may also be introduced to control for differences between the groups (which is important for balance at baseline as well as changes over time). Effect size estimates are included for DID analyses (see **Annex O, Exhibit O-1** for an example of DID analysis)

No matter what type of design is used, evaluators have to test for balance between the treatment and control/comparison groups at baseline, examining both key outcome variables and key predictors to ensure comparability between two groups. If groups are not comparable, evaluators must consider using matching techniques, such as propensity score matching, to improve the robustness of the design and analyses. If secondary data are available prior to collection of baseline data, these data can be used to select viable comparison schools, teachers, or students from which to collect baseline data. However, if such secondary data are not available, the evaluation team could consider increasing the sample size of at least the comparison unit to ensure good matches are available for each treatment unit, assuming adequate resources to do so. A balance table is included in all reports that use a QED and can be included in those that use an experimental design as well. This table is essential to show balance across measures and to alleviate some of the concern about selection bias.

Furthermore, for all types of designs, internal validity must be ensured via examinations of attrition, mortality, spillover, and history effects.

11.3 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.

11.4 Reporting Data Analysis

The purpose of analyzing EGRA data is twofold: 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 (see Annex O, **Exhibit O-2** for an example of graph used to visually report data).
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 (see Annex O, **Exhibit O-3**) 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 score distribution is 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 (CI) 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).

REQUIRED ANNEXES

Researchers must include details of the methodology and results of the analysis in the annexes, which can be quite lengthy and written in a technical language. The following annexes are to be included:

Details of the methodology, methods and data collection:

- Study objectives
- Design
- Data collection methods and process
- Data collection instruments
- Method and results of equating, if different tools were used at different study points
- Sampling parameters and attrition (for longitudinal studies)
- Details on weighting
- Limitations
- Results of test reliability analysis (Cronbach's alpha; item-total correlations)
- Intra-class correlation coefficient (ICC)

Details of analyses that are not included in the main report:

- Sample description
- Details of descriptive analyses
- Details of bivariate/multivariate analyses

12 USING RESULTS TO INFORM ACTION

12.1 A Strategy for Dissemination

The ultimate purpose of EGRA is to improve reading instruction and reading achievement. It is widely known that implementation of an assessment alone is not enough to attain this goal. Results have to be used in such a way as to inform policy, teaching practice, instructional support within and outside of school, and the use of resources to fill the right gaps in the system. Whether the solution comes in the form of training teachers to use better teaching methods, buying books to distribute to schools and classrooms, or mobilizing a community, the dialogue and actions that follow an EGRA are equally as important as collecting the data.

"The motivation for EGRA's creation was to gather timely access to information to inform learning improvement in low-income countries."

– Dubeck & Gove (2015)

Ensuring that results translate into action involves a multilayered approach beginning with planning and implementation efforts, which have been discussed elsewhere in this toolkit (for example, clearly defining the purpose of the assessment, carefully sampling the population to be assessed, including appropriate supplementary instruments such as classroom observations or questionnaires, and involving stakeholders in planning and implementation). Post-implementation analysis of the data focuses on actionable research questions through accurate analysis. Finally, results must be communicated to the appropriate audiences in a culturally and contextually suitable way in order to support understanding and action.

This section focuses on development of a strategic dissemination strategy, but urges readers to remember that the credibility—and subsequently effective use—of results will depend largely on careful execution of the previous steps in data collection.

Before planning results dissemination activities, implementers must reflect, at minimum, on the questions of "who?" "what?" and "how?"

1. **Who** will use the data?

And for each type of audience identified by the question above:

2. **What** data should be presented?

3. **What** type of information (and in what format) can the audience relate to best?

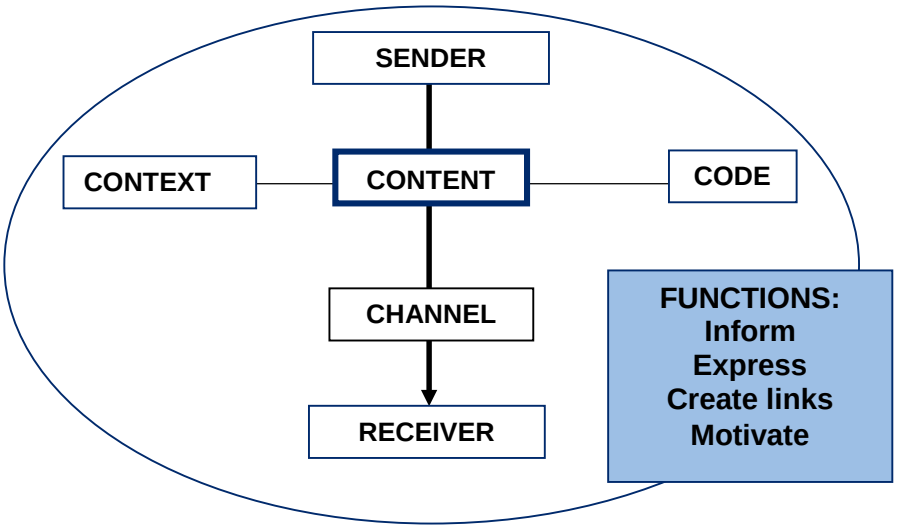
4. **How** will they use these data?

These questions will help shape the way results are prepared and shared among different stakeholders (e.g., communities, school-level officials, researchers, governments, ministry officials, teachers, parents). The EGRA implementing organization does not always need to disseminate the results widely if instead it can manage to reach the *right* people (at the right time, with the right message). When time or budget is limited, the focus is on reaching individuals who are influential and who have the capacity to connect with decision-makers and teachers.

12.1.1 Communicating Results

Exhibit 30 is a basic reminder of the elements involved in all types of communication. It is easy to focus only on the content of the message to be communicated; however, the clarity and impact of that message are shaped by the context in which it is situated, the "code" (language, tone) in which it is written, and the channel or means of communication (print, verbal, digital) with which it is sent. This means the communicator ("sender") must know the audience and be informed of how that audience is accustomed to accessing and processing information—including basic literacy skills and familiarity with data visualizations like charts and graphs. For some audiences, detailed technical information will be welcome, whereas for others the best approach may be to use the results to tell a story that paints a picture of what it means in context. Where the information is accessed and who disseminates the information can also change the way the message is received and interpreted.

Exhibit 30. A framework for communication



Source: Adapted from Jakobsen (1960).

Exhibit 31 from the *Guidance Notes for Planning and Implementing EGRA* (RTI International & International Rescue Committee, 2011, p. 82), provides an overview of potential audiences that can be targeted for disseminating EGRA results.

Exhibit 31. Overview of potential audiences

Level	Audience	Relevance
International	Donors	Donors can help support advocacy efforts, innovative pilot reading interventions, future assessments, and scaling up of best practices. Donor support is crucial, considering the limited resources that ministries of education in many developing countries have for innovation. Donors can help leverage EGRA findings, particularly when governments are unable or unwilling to take action.
	Academics or practitioners	Academics and other practitioners are often interested in EGRA survey findings because they provide valuable information about learning achievements and educational issues in many understudied contexts, and they help identify best practices.
National	Ministry or Department of education	The EGRA survey process and results can encourage government officials to do more to emphasize reading as a foundational skill. EGRA results can be a catalyst for a variety of government actions regarding early grade reading and also can increase interest in incorporating EGRA into national educational assessments. In some cases the government (or key elements within it) are already convinced of the need to act, and EGRA simply confirms, and increases the precision of, existing knowledge. EGRA can also shape the ensuing intervention.

Exhibit 31. Overview of potential audiences

Level	Audience	Relevance
	Budget authorities	EGRA results may help convince the government authorities that allocate funding to direct more public resources to the department/ministry of education for early grade reading. This typically will happen only when the financial authorities are convinced that the government and its partners have a viable strategy for remedying the situation EGRA detects. Thus, the idea that there are ways to improve the results may need to be part of the communication with these sorts of authorities.
	Teacher unions	Teacher support for early grade reading and related interventions is crucial. Outreach and collaboration with unions can influence teachers' perceptions of EGRA as a platform for positive change rather than as a means to criticize teacher performance. Teacher unions can help convey the importance of early grade reading, grade-appropriate expectations, and key findings to teachers as well as other audiences.
	Civil society and media	Civil society and media can help raise awareness, put pressure on decision makers (i.e., government) and, in some cases, promote sustainability.
Regional	Department or bureau heads	Provincial/state or district-level educational authorities are an important audience and partner, especially in cases where educational services are decentralized. Both competition and cooperation can be spurred on by EGRA results.
Community	Community leaders	Community leaders can help raise awareness about reading and good practices among community members—especially parents—and also exert pressure on local authorities and schools.
	Parents	Parents have a huge influence on children's reading habits and outreach. Interventions targeting parents can raise awareness about age-appropriate reading expectations, encourage reading in the home, and increase pressure on schools and policy makers to prioritize reading.
	Civil society	Civil society can support community-level dissemination and activities and also help increase accountability at the grassroots level.
School	Principals	Principals need to be aware of the importance of early grade reading and good practices in reading instruction in order to best support students, teachers, and parents. Principals can also be involved in interventions.
	Teachers	Teachers are a crucial audience to target for interventions and awareness-raising about early grade reading, age-appropriate expectations, and good practices for reading instruction.

Source: RTI International & International Rescue Committee, 2011, p. 82 [Table 6.2].

12.1.2 Dissemination Approaches

The results produced by EGRA or similar early grade assessment tools tend to concern policy makers and officials, especially when EGRA is used as a national,

system-level diagnostic or when the results are associated with an impact evaluation of a scalable (or scaled-up) instructional innovation. If it is true that political processes can "translate the will of people into public policies and establish rules that efficiently and effectively deliver services to all members of society" (Crouch & Winkler, 2008, p. 3) then addressing issues or challenges in an education system (an important service) begins with political will. A key assumption is data that show the learning outcomes of the education system can be used to stimulate that political will.

"In the long term, increasing country capacity to use information to inform instruction is critical to improving learning for the estimated 250 million children around the world who are not acquiring basic skills."

– UNESCO, 2014

National Level

It is best practice to validate results with the government (or other client/stakeholder) before further dissemination. This can be done through a one- or two-day event (often called a **"policy dialogue"**) that brings together either select government officials or broader representation from multiple stakeholder groups. The in-person format gives stakeholders an opportunity to ask questions and to provide country-specific background that can inform interpretation of the results and improve the final reporting. This type of workshop includes presentations of findings by the researchers; statements of policy and relevance from education officials or agencies implementing reading improvement programs; testimonies from field data collectors; and establishment of working groups to debate findings and action steps. This national workshop may be followed by regional or community meetings.

Of course, if the EGRA application was developed with the government in the first place, and there was a close relationship between the EGRA experts and the government experts, the validation will tend to go better. Thus, in most (but not all) cases, close collaboration from the beginning will reduce the likelihood of contentious dialogue during validation.

Apart from national and regional workshops, other dissemination strategies include preparation and dissemination of digital or printed reports; flyers, banners, and infographics; community meetings or multimedia (radio programs; video clips; documentaries). Using audio or video recordings during workshops, or as a dissemination strategy on their own, is an obvious and persuasive way to see the differences between a poor reader (a child reading at, say, 10–15 words per minute, with no comprehension) and a good reader (a child reading at, say, 60 words per minute, with comprehension). It is then much easier to put the quantitative results into this frame of reference.

CASE STUDY EXAMPLE: NATIONAL-LEVEL POLICY REFORM IN YEMEN

In early 2012, Yemen's Ministry of Education asked the Community Livelihoods Program (CLP), a USAID-funded development program implemented by Creative Associates International, to support the development of a new approach to teaching reading in the primary grades. The results of an EdData II-administered Early Grade Reading Assessment in 2011 in three governorates were presented in March 2012 in a policy dialogue with a broad range of stakeholders, including the Minister of Education. Creative Associates' CLP, based in Sana'a, facilitated the EGRA results review at a time when Yemen's early grade reading program was being developed.

The poor results of the 2011 EGRA (27% of grade 3 students could not read a single word of Arabic) strengthened the Ministry of Education's resolve to reform the way reading in Arabic is taught in the early grades. The Ministry prioritized early grade reading reform with a view toward setting a foundation for overturning years of underdevelopment in the education sector. The Minister of Education during this period took a central role in mobilizing the Ministry around the Yemen Early Grade Reading Approach (YEGRA). The USAID-supported YEGRA program was implemented as a trial in 310 schools during the 2012–2013 academic year, and as a mark of its success, the trial was then expanded to 800 schools in 2013–2014, and 1,200 schools in 2014–2015.

During the national dialogue in 2012, several factions wanted to put a stop to any curriculum changes until the constitution was finalized. This was an attempt to ensure that all parties would have an opportunity to be involved in any new curricula. The one exception to the moratorium on curriculum revision, as agreed upon by the delegates to the national dialogue, was the curriculum for grades 1–3, given that the YEGRA was already beginning to get attention for having made dramatic improvements in early grade reading, teacher skill development and motivation, and parental engagement.

The Minister of Education issued a number of decrees to ensure the success of the new YEGRA program. One decree was designed to ensure transparency and quality in the selection of trainers. Rather than having the Ministry appoint trainers from districts and governorates, a rigorous selection process was introduced. It included an application and selection process based on a set of criteria relevant for teaching reading of Arabic in the early grades.

Another decree issued was that in YEGRA schools, the mandatory time during each school day for reading in Arabic in grades 1–3 would be increased from 5 minutes to 70 minutes per day, five days per week. The Minister also issued a decree to ensure that the primary teachers who attended the YEGRA training were indeed the actual teachers of those grades. The Minister wanted to avert a common situation whereby favored teachers were selected by the principal to attend trainings regardless of their assignments to teach in grades and subjects not targeted by the training.

Ultimately, a policy decision to expand the YEGRA nationwide was made during the second trial year of the program, when the USAID-funded CLP was implementing YEGRA in more than 800 schools; the German aid agency GIZ was implementing in 72 schools; and with World Bank funding, the Ministry was implementing in 200 schools. The World Bank contributed support to the Ministry for nationwide expansion to 14,700 additional schools in 2014–2015. In other words, by the 2014–2015 school year, after the two years of the YEGRA trial, all 16,000 schools in the country were implementing the early grade reading program—including the 70-minutes-per-day standard—in all grade 1 classrooms.

Source: Adapted from du Plessis, El-Ashry, & Tietien (forthcoming).

Local Level

Citizens, particularly those who administer the EGRA test themselves (or simply ask children to read to them), quickly develop a sense that children are not reading and want to take action to be part of the change. Community members often seem to be taking notice of a serious reading problem among children in their schools. EGRA has helped induce this in some cases, but in other cases it is actually a response to concerns which have already been expressed. Most recently, early grade reading assessments have been used to garner community enthusiasm around education and other grassroots movements to develop awareness through evidence of students' reading (or not). The need for community mobilization and local awareness is an important "step toward increasing demand for education reforms that increase literacy" (Gove & Cvelich, 2011, p. 45).

To date, applications of EGRA have been used primarily to generate discussion at the national level and to spur ministries into action. EGRA results are intended to be reported for the lowest level or strata of the sample (often at the national level, but sometimes including regional or district levels). Because EGRAs are conducted on a sample basis, sharing school-specific EGRA results is not possible. Nor is it typically cost-feasible to conduct an EGRA for every school (and a large enough number of children in each school) to generate school-by-school results. However, because of the twofold benefit of raising community awareness and encouraging community engagement on issues of literacy and early grade reading, there are several strategies practitioners could consider.

The first recommended strategy is producing the EGRA findings in brief to share with school and community leaders for discussion points on the state of literacy generally (not specifically for the school but for the strata in which that school is located). This report is accompanied by explanations as to how each subtask relates to instruction and what teachers can do to improve student results. Sample lesson plans and suggested activities could also be shared with schools to indicate how school-

community stakeholders themselves could take action locally. Second, to obtain and report on school-level literacy scores, practitioners could use alternative literacy assessment tools such as those employed by Pratham for the Annual Status of Education Report (http://www.asercentre.org/p/141.html) as well as group-administered literacy assessment protocols developed through lot quality assurance sampling (LQAS) models (see the case study below as well as Batchelder, Betts, Mulcahy-Dunn, & Stern, 2015; Mulcahy-Dunn, Valadez, Cummiskey, & Hartwell, 2013; and Valadez, Mulcahy-Dunn, & Sam-Bossman, 2014).

CASE STUDY EXAMPLE: A PILOT AND CONTINUED PRACTICE OF LQAS MONITORING IN GHANA

Ghana established its National Inspectorate Board to develop tools to monitor the quality of education as part of its National Literacy Acceleration Programme. An LQAS pilot was conducted in Ghana as a way to test one such tool to monitor quality of education and identify areas that needed additional support at the local level. These pilot activities, which were designed to improve educational outcomes, included the use of EGRA to measure students' reading skills.

Results from the LQAS pilot study showed concern that students' reading scores were low across the sampled schools. However, the foundational reading skills, which are assessed by the EGRA, allowed the LQAS methodology to distinguish assessment results from one school to the next. The LQAS approach was meant to categorize districts between those that "performed at expectations" and those that "performed below expectations." These classifications were "based on whether 80% of schools achieved the traits of interest specified in a set of indicators related to teacher performance and pupil achievement" (Mulcahy-Dunn et al., 2013, p. 9).

After a policy dialogue session in Ghana about the LQAS pilot activities and the baseline EGRA, additional dissemination of the results occurred in the form of community-oriented "District Cluster Forums." These forums were used to disseminate the results more locally. It was during these forums that a need for continued monitoring at the district level was raised by local stakeholders. Based on this expressed interest and the success of the LQAS pilot at evaluating district and school performance in an efficient and cost-effective manner, LQAS monitoring received additional funding for expansion to several districts across Ghana.

12.2 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 fashion. 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.

CASE STUDY: SETTING NATIONAL BENCHMARKS IN KENYA

As of November 2015, USAID, through the EdData II project, had funded benchmarking and target-setting workshops in 12 countries: Egypt, Ghana, Jordan, Kenya, Liberia, Malawi, Mali, Pakistan, Philippines, Tanzania, West Bank, and Zambia. In each of these countries, early grade reading data were used to help draft benchmarks

From August 2011 to 2014, Kenya's Ministry of Education, Science and Technology (MoEST) implemented the Primary Math and Reading Initiative to improve fundamental skills in reading among students. PRIMR's design was inspired by an experimental reading improvement trial in the Malindi district in Kenya, carried out by the Aga Khan Foundation and RTI in 2007 (RTI International, 2008).

During PRIMR, the reading skills of randomly selected students from both treatment and control schools were measured through an EGRA. Because PRIMR was designed as a randomized controlled trial, it was feasible to determine the impact of PRIMR on learning. The data gathered from the EGRA were then used to inform ministerial policy decisions regarding investing in specific teaching methods that would lead to improved outcomes across gender and socioeconomic classifications.

Additionally, the MoEST invited PRIMR "to implement standardized research programs by working with the Kenya National Examinations Council (KNEC) in setting benchmarks for reading and numeracy" (RTI International, 2014b, p. 47). Results from PRIMR's baseline report helped gauge "appropriate benchmarks regarding student learning for fluency and comprehension" (RTI International, 2014b, p. 47). Furthermore, a Kenya Quality Education Meriting Tool was developed that could gauge early grade learning. PRIMR monitoring and evaluation staff developed the tool based on EGRA tools used in PRIMR. The MoEST then included the tool in its benchmark scoring levels. For the KNEC Steering Committee, PRIMR staff presented the research design, baseline findings and recommendations, and benchmark-setting results. During the latter presentation, PRIMR was able to show its level of precision of measuring student learning that influenced benchmarks. In an exercise during the KNEC meeting, Steering Committee members were asked to determine appropriate benchmarks for student fluency and comprehension using PRIMR baseline data gathered from the EGRA. Ultimately the Meriting Tool was modified to incorporate MoEST's benchmark scoring levels.

12.2.1 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.

As described previously (Section 12.1 on dissemination), 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. For comparison purposes, **Annex P** presents oral reading fluency norms for English.

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)

12.2.2 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.).

12.2.3 A Process for Setting Benchmarks

As mentioned in one of the case studies above, USAID's EdData II project had supported the drafting of benchmarks for reading performance in a dozen countries as of November 2015. In these countries, a consistent process was used to help identify acceptable levels of performance across several areas of reading skill development and grades. What follows are guidelines developed based on some of the lessons learned through the work in 12 countries.

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.

12.3 Cautions and Limitations

In some contexts, reactions to an EGRA-type reading assessment are not straightforward. Some commentators, in some countries, question the usefulness of oral reading fluency as a marker or precursor indicator of general learning, or even of reading. They may question why the assessment includes items, or formats, that do not directly reflect classroom instruction (for example, reading invented words). This is why it is important to have access to the background literature that explains the rationale, some of which is referenced in this toolkit; at the website of the International Literacy Association (www.reading.org); on the US National Institute for Child Health and Human Development's National Reading Panel pages (www.nationalreadingpanel.org); and on the website of the Center on Teaching and Learning of the University of Oregon, for the Dynamic Indicators of Basic Early Literacy Skills (DIBELS, http://dibels.uoregon.edu/).

In other cases, potential audiences seem to perceive that the EGRA efforts are trying to convey the notion that "reading is all that matters." In those cases, it is important to note that reading is indeed an important foundational skill that influences academic success across the school curriculum, and also that reading is a good marker for overall school quality. However, the effort is not based on the assumption that reading is all that matters.

In general, any attempt to measure education quality, as proxied by learning, is subject to these sorts of well-known debates. In the experience accumulating with the application of EGRA or EGRA-like tools, it seems that teachers, those concerned with direct support to teachers, and high-level officials see right away the value of EGRA, whereas some curricular or reading theoreticians have some trepidations or concerns with possible oversimplification. It is key to understand that the practical use of EGRA and the derived improvement strategies should be seen only as an entry point. The data can be used as an example of what can be achieved by focusing and monitoring specific results. The basic lesson can then be applied to other aspects of teaching and learning.

Resistance to the EGRA methodology and results may be strongest where results are weakest, which is why it is important to implement the assessment and analyze the results with rigor and objectivity. Additional contextual questionnaires (student, teacher, and school characteristics; classroom observations; etc.) can help explain performance outcomes, but administering them and analyzing the resulting data add costs to the implementation. When additional survey instruments are associated with EGRA results, implementers must pay careful attention to sample size and statistical significance, and avoid associating correlation with causation until further research has been done.

The Global Education First Initiative launched in September 2012 by the United Nations Secretary-General identifies the need for effective evaluation of student performance to improve education systems. Closely evaluating and monitoring how effectively a system is operating can lead to influence on policy, as it gives officials and decision-makers the opportunity to “use the information to direct support and resources for effective solutions (Office of the United Nations Secretary-General, 2012, p. 19).

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

Abadzi, H. (2012). *Developing cross-language metrics for reading fluency measurement: Some issues and options.* Global Partnership for Education working paper. Washington, DC: World Bank. Retrieved from https://www.academia.edu/3484052/Developing_Cross-Language_Metrics_for_Reading_Fluency_Measurement_Some_issues_and_options._World_Bank_Global_Partnership_for_Education_working_paper

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, US Department of Education; and the US 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'Education (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

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. 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

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

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., & Ialongo, 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: Guildford.

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

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.

Kandhadai, P., & Sproat, R. (2010). Impact of spatial ordering of graphemes in alphasyllabic scripts on phonemic awareness in Indic languages. *Writing Systems Research*, *2*(2), 105–116.

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_AKanjee_Eng.pdf

Karanth, P. (2002). Reading into reading research through nonalphabetic lenses: Evidence from the Indian languages. *Topics in Language Disorders, 22*(5), 20–31.

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

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-k/borer/Memoire/incidentiallearning/incidentiallearning.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

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

National Institute of Child Health and Human Development (NICHD) [US]. (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

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

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. Retrieved from 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. (2003). The universal grammar of reading. *Scientific Studies of Reading, 7*(1), 3–24.

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

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

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.

Pouezevara, 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

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

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

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

RTI International. (2015). *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 & 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 supralexical 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.

Stern, J. & Nordstrum, L. (2014). *Indonesia 2014: The National Early Grade Reading Assessment (EGRA) and Snapshot of School Management Effectiveness (SSME) survey.* Prepared for USAID/Indonesia under the Education Data for Decision Making (EdData II) project, Task Order No. AID-497-BC-13-00009 (RTI Task 23). Research Triangle Park, NC: RTI International. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=680

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

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

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

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. Retrieved from 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

World Bank. (2015a). *EdStats dashboards: Learning outcomes dashboard* [Web page]. Washington, DC: World Bank. 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-Dağli, 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

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: INFORMATION ABOUT 2015 EGRA WORKSHOPS

Source: EdData II project website, News and Events, https://www.eddataglobal.org/news/index.cfm

A.1 Workshop on Designing and Implementing EGRA: Understanding the Basics

24 Mar 2015 – Kellie Betts

During March 2–4, 2015, RTI technical staff facilitated the workshop "Designing and Implementing Early Grade Reading Assessments: Understanding the Basics." The workshop was hosted by the Global Reading Network at University Research Co. (URC), LLC, in Bethesda, Maryland. Kate Batchelder, Alison Pflepsen, and Sarah Pouezevara led the EGRA workshop, designed to teach both in-room and online participants the basics of designing and implementing early grade reading assessments.

The foundation for the overall curriculum was the EGRA Toolkit and Guidance Notes. Field-based knowledge and practice were intertwined in several of the sessions. Richard Vormarwor of Educational Assessment and Research Center (EARC) in Ghana and Eva Yusuf of Myriad Research in Indonesia—both RTI subcontractors—shared specific stories and experiences from their respective contexts.

The opening session was led by RTI literacy expert Margaret "Peggy" Dubeck and RTI Director of Research Amber Gove, who gave a history and overview of the EGRA instrument. Throughout the workshop, participants received guidance, learned about best practices, and participated in interactive practice on several different aspects of EGRA design and implementation. Among the topics covered were research design and sampling framework; adaptation; administration, scoring, and data capture; electronic data collection; training of data collectors; data collector assessment and selection; pilot and data collection; and data dissemination.

Elena Vinogradova of Education Development Center (EDC) delivered a brief presentation on electronic data capture, after which the participants received hands-on practice with tools such as "SurveyToGo" and Tangerine®. Ben Sylla and Christie Vilsak of USAID also gave guest presentations as part of the workshop. The wrap-up session, "Implementation: Practice Makes

Perfect," allowed participants to bring together all the materials and information from the previous sessions to begin planning and designing an early grade reading assessment.

The 3-day workshop was funded by the USAID/Washington EdData II task order called Measurement and Research Support to Education Strategy Goal 1.

A.2 Technical Staff Contribute to Improving the Quality of EGRA Data

16 Jun 2015 – Kellie Betts

The Global Reading Network hosted two days of panel discussions on advanced topics related to early grade reading assessment (EGRA) design, administration, analysis, and reporting. RTI International technical staff presented on each of the topics alongside additional expert panelists from various organizations. The event was held at University Research Co., LLC (URC), May 27–28, 2015.

The event was funded by the United States Agency for International Development (USAID), which is in the process of developing enhanced guidelines for EGRA design, administration, and reporting. This series of presentations and discussions was designed to bring together organizations and members of the Global Reading Network to share expertise and experiences. Thirty participants attended the event in person, and dozens more from various EGRA-implementing organizations participated online via WebEx Webinar. The participants had extensive (current or prior) experience in EGRA planning, administration, and/or reporting, allowing them to contribute to the technical presentations and rich discussions that defined the event.

Each session lasted two hours and included presentations from expert panelists as well as facilitated discussions among the panelists and participants. Each discussion focused on the topic from the preceding presentation and typically consisted of clarification questions, suggestions, and other general comments. At the conclusion of each discussion, the facilitator finished the session by highlighting areas of consensus among panelists and participants or areas that would require further discussion and attention.

To open the event on Wednesday, May 27, URC and USAID staff welcomed attendees and introduced the presenters. The first session covered research design and sampling frameworks. RTI's Chris Cummiskey presented alongside Matt Sloan (Mathematica Policy Research, Inc.) and Elena Vinogradova (Education Development Center [EDC]). The facilitated discussion was led by Melissa Chiappetta of Social Impact.

The second session, on creating comparable EGRA instruments across languages, included expert panelist presentations from Margaret “Peggy” Dubeck from RTI, Carol de Silva from FHI 360, and Fathi El Ashry from Creative Associates. Pooja Reddy Nakamura (American Institutes for Research [AIR]) led the discussion on this topic.

The final session on the event’s first day focused on inter-rater reliability. RTI’s Simon King, Jeff Davis of Management Systems International (MSI), and Abdullah Ferdous of AIR presented. Fathi El Ashry facilitated the subsequent discussion among the panelists and participants.

The event’s second day, Thursday, May 28, began with a session on EGRA/early grade math assessment (EGMA) data preparation and analysis, facilitated by Agaia Zafeirakou from Global Partnership for Education (GPE). Panelists for this session included Simon King (RTI), Elena Vinagradova (EDC), and Melissa Chiappetta (Social Impact).

The second session, on equating EGRA across same-language applications, was presented by Jonathan Stern of RTI, Jeff Davis of MSI, and Zarko Vukmirovic of AIR. Alla Berezner of Australian Council for Educational Research led the discussion that followed.

Jeff Davis and Thomaz Alvares (MSI) gave an abbreviated presentation to discuss the Goal 1 Education Strategy and the current proposed methodology refinements. Following the presentation, USAID’s Benjamin Sylla fielded questions and answers between the audience and panelists.

The final session of the second day, facilitated by Jill Meekes (Chemonics), included three presentations on how to create public-use files for data sets. Chris Cummiskey and Kellie Betts (RTI), Thomaz Alvares (MSI), and Roger Stanton (Optimal Solutions) presented on the topic.

Both days of panel sessions fostered technical discussions and raised issues for further consideration regarding the various topics of EGRA administration. The participating panelists plan to work together to make recommendations and develop guidelines on each of the topics that were presented. The event was highly praised, with a large majority of online and in-room participants agreeing that the content from the sessions was informative, encouraged participation and interaction, and led to engaging discussions.

ANNEX B: SAMPLE SIZE CONSIDERATIONS IN EARLY GRADE READING ASSESSMENTS

B.1 Introduction

This annex sets out basic sample size considerations applicable to Early Grade Reading Assessment samples. It is designed to inform Ministry or Department of Education staff, donors, or other actors interested in setting up an EGRA on sampling size requirements and calculations.

B.2 Sampling Approach

The applied sampling approach will impact the sample size requirements. Other things being equal, selecting students randomly from a national listing will require a smaller sample size, whereas *clustered samples* will require relatively larger sample sizes. Although it may appear contradictory, pure simple random samples are relatively expensive when compared to other sampling methods. If one tried, for example, to apply a pure simple random sample of 400 children, one might be faced with a situation of having to go to nearly 400 schools, and then test only one child in each school, which would increase transportation and labor costs tremendously.[29]

In addition, one would in principle need a list of all the schoolchildren in the country, and their location, to obtain a simple random sample of children. Such lists simply do not exist in most countries. With school-based sample clustering, schools are selected first, and then students within schools (clusters) are selected. Randomly sampling schools first, and then children, reduces travel costs and travel time and it eliminates the need to rely on a national listing of students. Since much of the cost of surveys is getting to the schools in the first place, one may as well test as many children as it is feasible to test in each school in a one-day visit, as a way of

[29] There would be a need to go only to *nearly* 400 schools because, by luck of the draw, and depending on the total number of schools in the country, some schools would have more than one child selected. In a country with, say, only 500 schools, sampling 400 children via a simple random sample is quite likely to yield several cases where there is more than one child per school, whereas this would not be the case in a country with, say, 80,000 schools.

increasing sample size at relatively low cost. (Refer to Annex C of this toolkit for more about complex and cluster sampling.)

Past EGRA applications have shown that it is possible for one assessor to interview between 4 and 10 children in one school day, depending the number of subtasks and questions asked to each student.[30] Assuming, *as an example only,* a sample of 15 children per school, a sample size of 400 would require one to visit only some 27 schools—a considerable economy over having to visit 400 or so. (The actual desired sample of children per school may vary based on country characteristics.) Therefore, a cluster sampling approach is recommended.

However, applying the cluster approach results in a loss of realism because children typically vary less within schools than the "representative child" in each school varies from children in other schools. Intraclass correlation (ICC) comes into play, in that children within schools tend to belong to the same social class, or have the same language advantage or disadvantage, or have similar quality of teachers and be exposed to similar management practices as each other—to a greater degree than children in different schools. In this sense, the true or population variability between children tends to be underestimated if one uses a cluster sampling approach—that is, the transportation and labor cost efficiency is gained at the price of a loss of information about variability and hence, unless adjustments are made, there will be a loss in precision. Fortunately, there is a measurement that will tell us the degree to which the clustering may be leading to an underestimate of variability. This measure, known as the *design effect (DEFF)*, can be used to adjust the sample size to account for the loss in variability caused by clustering.

Four items need to be included in our sample size calculation. These are:

1. *Variability* in student reading scores (or other EGRA variable if desired) both overall variability and variability within schools, and across schools
2. Researcher-determined *confidence interval (CI) width*
3. Researcher-determined *confidence level* (typically 95%)
4. *Design effect (DEFF)* caused by the application of cluster sampling.

[30] This specific number of children that can be interviewed depends on the version of the EGRA instrument being applied, the number of languages in which the EGRA is being carried out, and whether the EGRA is part of other research taking place at the school.

B.3 Calculating Sample Size for a Given Confidence Interval and Confidence Level

Formulaically, the needed sample size may be represented as follows:

$$n = 4\left(\frac{CLtvalue\ DEFT\ SD}{CI_{Width}}\right)^2,$$

where:

n is the sample size needed;

CLtvalue is the *t*-value associated with the selected confidence level (typically 1.96 for 95%),

DEFT is the square root of the design effect (DEFF),

SD is the estimated standard deviation, which is a measurement of the variability in the chosen variable;

CI_{Width} = the researcher-determined *width of the confidence interval;* and

the number 4 is derived from the basic equation for a confidence interval.[31]

As may be seen from this equation, increases in the *confidence level,* the *design effect,* and the *variability* (as measured by the SD) all work to increase the required sample size *(n)*. Any increase in the *Width* of the confidence interval*,* conversely, reduces the sample size requirement but it also reduces precision, by definition.

For purposes of developing sample size recommendations, the square root of the design effect (DEFT being square root of DEFF) and the standard deviation (SD) are calculated using data from some previous EGRA applications.

The DEFF is calculated as follows:

---

[31] This equation is derived from the traditional formula for a confidence interval as $\overline{X} \pm CLtvalue\frac{SD\ DEFT}{\sqrt{n}}$, where the expression on the right of the ± sign is the one-sided width. The total two-sided width then is $Width = 2\ CLtvalue\frac{SD\ DEFT}{\sqrt{n}}$. Algebraic manipulation will then get one to the equation used in the main text and will show why the 2 becomes a 4.

$$DEFF = 1 + (clustersize - 1)\, ICC\ ,$$

where:

clustersize is the size of the average cluster (the number of children sampled in each school[32]), and

ICC is the intraclass correlation coefficient.

Increases in *clustersize* or in the ICC will increase the design effect. If we were to sample students using simple random sampling, then the *clustersize* would be 1 (one child per school in the sample), the ICC would be zero because there was no other sampled student in the school to compare, and the DEFF would be 1. That is, clustering does not affect estimated variability if *clustersize* is only 1.

The ICC is a measure of how much of the variability lies between schools and how much lies within schools. An intuitive way to think about it is that it indicates the probability of finding two observations that are the same in the cluster relative to finding two identical *randomly* selected observations. For example, an ICC of 0.41 would indicate that one is 41% more likely to find two students with the same reading fluency within a cluster (school) than one is to find two students with the same fluency levels pulled at random out of any two schools.

There are various understandings of the ICC in the literature. The ICC in this context follows the usage in Stata software, and is calculated as follows:

$$ICC = \frac{MSE_{between} - MSE_{within}}{MSE_{between} + (clustersize - 1)\, MSE_{within}}\ ,$$

where:

MSE is the mean squared error, and

clustersize is the average size of clusters (the number of children in each selected school).

$MSE_{between}$ measures the amount of variation that exists between schools (clusters). Arithmetically, $MSE_{between}$ is the sum of squared deviations between each cluster's (school's) mean and the grand mean, weighted by the size of the cluster (the number

[32] Assuming that schools are the first item sampled. In some surveys, geographical areas are sampled first.

of children sampled in the school). MSE_{within} measures the amount of variation that exists within schools (our clusters). Arithmetically, MSE_{within} is the sum of the squared deviations between each child and the cluster (school) mean, divided by the total number of children minus the number of clusters. In symbols,

$$MSE_{between} = \frac{\sum_{j=1}^{cluster} n_j (\overline{X}_j - \tilde{X})^2}{cluster - 1}$$

and

$$MSE_{within} = \frac{\sum_{j=1}^{cluster} \sum_{i \in j=1}^{n_j} (x_{ij} - \overline{X}_j)^2}{\sum_{j=1}^{cluster} n_j - cluster},$$

where:

$\tilde{X}$ is the "grand" or overall mean,

j is an index for clusters,

$i \in j$ is an index for the *i*th child in cluster *j,*

$\overline{X}_j$ is the mean of the *j*th cluster (or school),

cluster is the number of clusters or the index of the last cluster, and

n_j is the size of the *j*th cluster or the index of the last member of the *j*th cluster.

The analysis of variance (ANOVA) procedure in Excel may be used to calculate both MSE_{within} and $MSE_{between}$.

Exhibit B-1 shows a range of estimates of both the ICC and the DEFT for a few particular cases and the implication of these variables for the number of schools (clusters) and resulting total sample size. An SD of 29 is assumed for all cases for demonstration purposes, a total confidence interval width (two-sided width) of 10 is specified, and a confidence level of 95% is used. The ICC, DEFT, and *clustersize* are actual values from EGRA studies.

Exhibit B-1. Estimated *ICC* and *DEFT* across a variety of countries and grades, showing the average cluster size in each case

Country	ICC	DEFT	*clustersize*	*n*
Country A, Grade 3	0.17	1.2	3.75	198
Country B, Grade 2	0.22	2.3	20	698
Country C, Grade 3	0.25	1.6	7.57	356
Country D, Grade 3	0.47	2.3	10.05	708
Country E, Grade 2	0.48	1.8	5.35	416

Source: Calculated by RTI International from various EGRA surveys.

The DEFTs in Exhibit B-1 are affected by the ICC and also by the cluster size. As can be seen in the equation for the DEFT, both affect the DEFT. In Country B, for example, the DEFT turns out to be a little high (2.3), even though the ICC is low (0.22), because the cluster size is 20; so one suppresses a lot of variation by taking so many of the children from specific schools. In Country D, the driver behind the high DEFT is the high ICC. In Country A, the DEFT is the lowest because both the cluster size and the ICC were low. The impacts on required sample size are significant. In Country A, a sample of only 198 children would be needed (but some 53 schools), whereas in Country D, a sample of 708 children and 70 schools or so would be needed.

B.4 Recommended Sample Sizes for Confidence Intervals

In determining actual recommended sample sizes, a reasonable requirement would be that differences between grades are "meaningful" in some sense—e.g., the overall confidence intervals are sufficiently narrow that the confidence intervals for contiguous grades do not overlap. If we know that the average inter-grade difference is 14, a *Width* of 14 is sensible.

If one assumes a Width of 14, an ICC of 0.45, a cluster size of 12, and an SD of 29 the "right" sample size is 409 children. Note that as the grade increases, so do the student scores; as a result the standard deviation almost certainly increases as well. As a result, higher grades generally need larger sample sizes to achieve the same level of precision.

Given the typically very small differences between boys' and girls' average performance on EGRA subtasks (and/or given that the gender differences are highly variable across countries, unlike the steady grade progression), and given the equation for sample size, it must be clear that a very small *Width* would be needed to

detect gender differences, and hence a very large sample size: around 7,000. It seems wise to accept the notion that most reasonable sample sizes are not likely to capture statistically significant differences between boys and girls. This highlights, in passing, the importance of distinguishing between *substantive difference* and *statistically significant difference.* In general, if there is any difference at all between any two subpopulations, even if it is not substantively interesting, researchers could "force" it to become statistically significant by drawing an enormous sample. In other words, small differences that are of marginal interest may be determined to be statistically significant by a large sample size. The judgment being made here is that gender differences on the EGRA have tended to be sufficiently small that only very large samples can detect them with statistical significance.

As of late 2015, Early Grade Reading Assessments have been conducted in many countries. When further assessments are being conducted in countries where assessments have already been completed, we can use the public-use file (PUF) data from the first assessment to make more precise estimates for the new assessment. PUF data from Education Data for Decision Making (EdData II) task orders can be requested from the EdData II website (https://www.eddataglobal.org/datafiles/index.cfm?fuseaction=datafilesIndex). For example, extracting the ICC and standard deviation from the data set from an EGRA conducted in Zambia in 2012 made possible a more precise estimate for sample size for the 2014 National EGRA in Zambia. Within Stata, the command *loneway* can be used to determine the ICC and the command *summarize* can be used to determine the standard deviation.

Consideration must be given to potential differences between the intended populations of interest for the previous and future EGRAs. If different languages, regions, or grades are being assessed, the ICC and standard deviation from the previous EGRA might not be accurate for future assessments. However, if these hurdles can be overcome, using historical data presents an opportunity to calculate reasonable sample size estimates. **Exhibit B-2** shows how to vary the number of students sampled from 10 to 22 (in 2-point increments) and vary the confidence interval widths between 10, 12, and 14. Using fixed values of 26 for the standard deviation and 0.45 for the ICC, we can then calculate the DEFT and thus estimate the number of students and schools required. As shown in Exhibit B-2, there is little benefit in having more students per school and fewer schools, because the ICC is high; 10 students per school and 52 schools provide the same level of precision as 22 students per school and 49 schools, a difference of 561 students (1,086 minus 525). So, by assessing many more students, we can go to just three fewer schools—for which there is almost certainly no financial benefit.

Exhibit B-2. Estimated number of students and schools required based on varying the number of students per school and confidence interval width and keeping the ICC and standard deviation fixed

Values to vary		Fixed values			Outcomes	
Number of sampled students in school	95% confidence interval width	Standard deviation	ICC	DEFT	Total number of students	Number of schools
10	±5	26	0.45	2.25	525	52
12	±5	26	0.45	2.44	618	52
14	±5	26	0.45	2.62	712	51
16	±5	26	0.45	2.78	805	50
18	±5	26	0.45	2.94	899	50
20	±5	26	0.45	3.09	992	50
22	±5	26	0.45	3.23	1,086	49
10	±6	26	0.45	2.25	364	36
12	±6	26	0.45	2.44	429	36
14	±6	26	0.45	2.62	494	35
16	±6	26	0.45	2.78	559	35
18	±6	26	0.45	2.94	624	35
20	±6	26	0.45	3.09	689	34
22	±6	26	0.45	3.23	754	34
10	±7	26	0.45	2.25	268	27
12	±7	26	0.45	2.44	315	26
14	±7	26	0.45	2.62	363	26
16	±7	26	0.45	2.78	411	26
18	±7	26	0.45	2.94	458	25
20	±7	26	0.45	3.09	506	25
22	±7	26	0.45	3.23	554	25

B.5 Hypothesis Testing Versus Confidence Intervals: Sampling Implications

In deciding about sample sizes, one factor to be taken into account is whether the bases for comparisons between groups (e.g., between fluency levels in different grades) are non-overlapping confidence intervals or one-sided hypothesis tests. A common practice is to present CIs for key variables, and to state or imply that non-overlapping CIs are a useful first cut at seeing whether differences between groups are significant. This is often done because the researcher does not know ahead of

time what contrasts, or hypothesis tests, will be of most interest. In that sense, presenting CIs for key variables, in EGRA, seems like a wise practice. In addition, in general, readers with a substantive interest in the matter care a great deal about the actual parameters being estimated (the mean levels of fluency, for example), and their likely range, and might care less about whether differences between subpopulations of interest are statistically significant.

However, trying to make CIs narrow enough not to overlap, and hence detect a given difference between means, requires larger sample sizes. Doing one-sided hypothesis tests might require smaller sample sizes. On the other hand, hypothesis tests are harder to interpret, drawing attention perhaps overmuch toward "statistical significance" and somewhat away from the parameters under consideration. Furthermore, some of the economy in doing hypothesis tests can be achieved only if the hypothesis tests are one-sided.

There is some debate in the evaluation literature on the conditions that justify one-sided hypothesis testing. The debate is not conclusive, however, so it may be useful to recall the issues at hand.

Hypothesis testing generally posits a "null" hypothesis that, say (using fluency as an example), the fluency level for a given grade is equal to the fluency level for a previous grade, or that the fluency level after an intervention is the same as the fluency level before an intervention. Then one posits alternative hypotheses. One form of an alternative hypothesis is that the fluency level in a higher grade is simply different from the fluency level of a previous grade, or that the fluency level after an intervention is different from the fluency level before the intervention. To test this hypothesis, one then carries out a "two-sided" hypothesis test. This is common when one is interested in exploratory analyses, where a certain treatment or variable (level of rurality, experience of the teacher, etc.) might have either a positive or negative impact on something else (test scores might be impacted negatively or positively by degree of rurality, and one does not have a strong *a priori* reason to test a hypothesis going in a particular direction).

In most EGRA applications, it seems reasonable to believe that most of the hypotheses being tested, or most of the statements one might wish to make, are uni-directional. Thus, one might be justified in positing one-sided hypothesis testing, to achieve economies in sample size. If there are good reasons to believe the analysis needs to be more exploratory and descriptive in nature, then two-sided hypothesis testing is used.

If desired, confidence intervals can be presented along with hypothesis tests. The purpose of presenting the CIs is to foster a focus on the parameter in question, such as oral fluency in reading connected text. But it has to be noted that if sample sizes

are just large enough to allow detection of differences in one-sided hypothesis tests, the CIs will tend to be relatively wide. Thus, the EGRA approach decides first whether one-sided hypothesis tests are acceptable, with the proviso that this might mean slightly wider CIs. The following discussion highlights the issues.

Suppose we have two sample means, $\overline{X}_1$ and $\overline{X}_2$. To keep things simple, let us say that the estimated standard errors (SEs) for both are the same, so $SE_1 = SE_2 = SE$. We also assume, without much loss of generality, that this is due to equal SDs and equal sample sizes.[33] For this discussion we will stay with 5% tests or 95% CIs. The *t* ordinates are assumed to be for the appropriate degrees of freedom. The 95% CIs are

$$\overline{X}_1 \pm t_{.025}\ SE$$

$$\overline{X}_2 \pm t_{.025}\ SE\ ,$$

where $t_{.025}$ is the *t* ordinate required for a two-sided 5% test with the appropriate degrees of freedom. The requirement that the two CIs for each mean not overlap is equivalent to requiring that

$$\overline{X}_1 + t_{.025}SE < \overline{X}_2 - t_{.025}SE$$

or

$$\overline{X}_2 - \overline{X}_1 > t_{.025}SE + t_{.025}SE = 2t_{.025}SE$$

if the first estimated mean is smaller than the second one, and similarly, but with different signs, if the second is smaller; or more generally:

$$\left|\overline{X}_1 - \overline{X}\right|_2 > 2t_{.025}\ SE\ ,$$

because the CIs for the means are symmetrical around the mean, and have the

[33] In fact, most of the SDs and SEs will differ from each other. Sample size and SD equality are assumed in *this* exposition solely for the sake of simplicity.

same width, given that the SEs and degrees of freedom (as driven by *n*) are the same.

But the requirement that the CI for the *difference* not overlap *with 0* is equivalent to requiring that

$$\left|\overline{X}_1 - \overline{X}_2\right| > 1.41\, t_{.025}\, SE\ ,$$

because of the equation for the standard deviation for a difference between means, which is as follows, given the assumption of equal standard deviations and equal samples:

$$SD_{diff} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}} = \sqrt{2\frac{SD^2}{n}} = 1.41 SD\ .$$

Note that the ratio of 2 to 1.41 is 1.41, as any number divided by its square root is equal to its square root. This means that in the first case, one would need a smaller SE than in the second case, so as to create no overlap of the CIs—smaller by 1.41 times. Given that $SE = SD/\sqrt{n}$, an *SE* that is 1.41 times smaller requires a sample size that is 2 times bigger, as

$$\frac{SE}{1.41} = \frac{SD}{1.41\sqrt{n}} = \frac{SD}{\sqrt{2n}}\ .$$

The following instant tests from Stata (using the "ttesti" command) serve to illustrate. The tests use the values already used in the illustrations above. For the sake of illustration of the basic principle regarding the differences between confidence intervals and hypothesis tests, we focus on a case where the DEFF is 1. The *procedure* used is that for unequal variances, although in practice and to make the exposition easier, the standard deviations input into the illustrations are equal to each other. Note that the ttesti command cannot be used for most EGRA analysis because it does not adjust the standard errors to account for the clustered sample design.

First, we have a case where the confidence interval for the *difference* between the two means does not overlap zero, but almost does, as noted in the lower highlighted area. Notice that Stata presents the CIs for each variable, the CI for the difference

between the variables, and all relevant hypothesis tests for the difference between the variables.

```
ttesti 34 20 29 34 34 29, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      34          20    4.973459          29    9.881422    30.11858
       y |      34          34    4.973459          29    23.88142    44.11858
---------+--------------------------------------------------------------------
combined |      68          27    3.593661    29.63409    19.82702    34.17298
---------+--------------------------------------------------------------------
    diff |                 -14    7.033533               -28.0429     .042902
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  -1.9905
Ho: diff = 0                     Satterthwaite's degrees of freedom =       66

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0253         Pr(|T| > |t|) = 0.0507          Pr(T > t) = 0.9747
```

The CIs for both means overlap considerably, as noted in the two upper highlighted areas, but the CI for the *difference* does not overlap zero (though it almost does, by design) as can be noted in the lower highlighted area. Yet, this is really the correct way to interpret the requirement of detecting a difference between the groups. To avoid the overlap in the CIs for the means themselves, one would have to double the sample sizes.

The following test shows that with a doubling of the sample size, the CIs for the individual means just barely miss overlapping, as shown in the upper highlighted areas:

```
ttesti 69 20 29 69 34 29, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      69          20     3.49119          29    13.03344    26.96656
       y |      69          34     3.49119          29    27.03344    40.96656
---------+--------------------------------------------------------------------
combined |     138          27    2.531281    29.73582    21.99457    32.00543
---------+--------------------------------------------------------------------
    diff |                 -14    4.937288               -23.76379   -4.236213
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  -2.8356
Ho: diff = 0                     Satterthwaite's degrees of freedom =      136

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0026         Pr(|T| > |t|) = 0.0053          Pr(T > t) = 0.9974
```

But a doubling of sample size is a high (and unnecessary) price to pay to have non-overlapping CIs for the means, rather than a non-overlapping-with-zero CI for the difference between the means. This can be seen by the fact that the CI for the difference between the means is quite far from zero (middle highlight), or by the fact that a two-sided hypothesis test for the difference between the two means yields a probability value way below the 5% threshold (lowest highlight).

Yet one has even a little more leeway. Most of the gain in efficiency between hypothesis testing over the notion of "non-overlapping confidence intervals" is achieved simply by posing the problem as a hypothesis test. But, if desired and if justified *a priori,* a little more efficiency can be gained by supposing a one-sided hypothesis test. Note that in the first Stata printout above, even though the CI for the difference almost touches zero, a *one-sided* hypothesis test is very strong—"overly" strong relative to a 5% test. Because the 95% CI for the difference almost touches zero, the probability value for a *two-sided* hypothesis test is indeed 0.05 (or close to it), as one would expect given the equivalence between a two-sided hypothesis test and a CI for a difference between means that does not include zero. But the probability value for a one-sided hypothesis test, in the first run above, is only 0.025 (0.0249 actually), so we have degrees of freedom to spare if all we want is a 5% test. Since the *t* value for a one-sided 5% hypothesis test is 1.67 (or thereabouts, for large *n*), whereas that needed for a two-sided one is around 1.96, we could make the sample smaller by a ratio of approximately $\sqrt{1.67/1.96} = 0.73$.

In effect, we are requiring only that

$$\left|\overline{X}_1 - \overline{X}_2\right| > 1.41\, t_{.05}\, SE$$

for a one-sided t-test, with $t \approx 1.67$ with any reasonably high *n*.

The following instant Stata test demonstrates that when the sample size is reduced, from the first set of results, to a ratio of 0.73 of 34, or 25, then the one-sided hypothesis test has a probability value just under 0.05, as needed (lower highlight). The CIs now totally overlap (upper highlights). The 95% CI for the difference even overlaps with zero, because requiring a non-overlapping-with-zero CI for the difference would be equivalent to a two-sided hypothesis test.

```
ttesti 25 20 29 25 34 29, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      25          20         5.8          29    8.029388    31.97061
       y |      25          34         5.8          29    22.02939    45.97061
---------+--------------------------------------------------------------------
combined |      50          27    4.180518    29.56073    18.59893    35.40107
---------+--------------------------------------------------------------------
    diff |                 -14    8.202439               -30.49211    2.492108
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  -1.7068
Ho: diff = 0                     Satterthwaite's degrees of freedom =       48

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0472         Pr(|T| > |t|) = 0.0943          Pr(T > t) = 0.9528
```

Taking both factors together, the sample size needed for a one-sided hypothesis test is about 0.36 of what is needed to create non-overlapping (two-sided) CIs on the two means.

Note that if the SD is effectively augmented by a DEFT of 2.44 (the result of the same assumptions as were used in establishing the sample size of 409 for a CI, namely an ICC of 0.45 and a cluster size of 12), then the sample size needed for a 5% test goes up, essentially up to 2.44^2 times 25, or 148.

```
ttesti 148 20 70.7 148 34 70.7, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |     148          20    5.811504        70.7    8.515112    31.48489
       y |     148          34    5.811504        70.7    22.51511    45.48489
---------+--------------------------------------------------------------------
combined |     296          27    4.122578    70.92751    18.88661    35.11339
---------+--------------------------------------------------------------------
    diff |                 -14    8.218708               -30.17496    2.174957
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  -1.7034
Ho: diff = 0                     Satterthwaite's degrees of freedom =      294

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0448         Pr(|T| > |t|) = 0.0895          Pr(T > t) = 0.9552
```

These factors allow some economy in sample size with a one-sided hypothesis test as opposed to non-overlapping confidence intervals. However, there is an opposite pressure, namely the need to take power into account. Taking power into account, assuming a power of 0.8 and a 5% hypothesis test, and introducing the notion that SDs *might* be different, a sample size for a one-sided hypothesis test is

$$n = \frac{(SD_1^2\ DEFF + SD_2^2\ DEFF)(0.85 + 1.67)}{DIFF^2},$$

where:

0.85 is the one-sided *t* value for a power of 0.8,

1.67 is the one-sided *t* value for a 5% test (both with 60 degrees of freedom, an appropriately low number), and

DIFF is the hypothesized difference between, say, grades.

Using the same parameters as for the confidence interval, namely a DEFF of 5.595 (DEFT of 2.44) (due to an ICC of 0.45 and a cluster size of 12), and SDs of 29 (meaning that for this example they happen to be the same, but using the equation that allows for different SDs), and a DIFF of 14, the required sample size is 324. In the more pessimistic case where the SDs are 50, but the DIFF is allowed to be 20, the sample size needed is 472. In either case these are a little smaller than what is needed for a 95% confidence interval.

If one were to conclude, based on the sorts of discussions above, that two-sided tests were more appropriate, then the correct equation would be:

$$n = \frac{(SD_1^2\ DEFF + SD_2^2\ DEFF)(0.85 + 2)}{DIFF^2}.$$

In that case, and using the same assumptions as above, the sample size with an SD of 29 is 414, and with the more pessimistic SD of 50 but a DIFF of 20, it would be 603.

B.6 Summary of Sample Sizes Based on Confidence Intervals and Hypothesis Tests

Exhibit B-3 summarizes a range of suggestions on sample sizes. However, if historical EGRA data are available for that country, that data are used as a priority to estimate more appropriate ICC and standard deviations. The exhibit assumes an SD of 29, an ICC of 0.45 (which is on the high end of what typically has been found in EGRA studies), and *clustersize* (number of sampled children per school) of 10. In the case of hypothesis tests, a power of 0.8 is assumed. In each case, the number of schools needed is derived by rounding up the result of dividing the sample size by 10.

Exhibit B-3. Summary of sample sizes according to various considerations

	Sample size	No. of schools
Confidence level 90%		
Confidence interval approach:		
Two-sided width of interval: 10	475	48
Two-sided width of interval: 15	211	22
Hypothesis testing approach – one-sided:		
Minimum detectable difference: 10	390	39
Minimum detectable difference: 15	173	18
Hypothesis testing approach – two-sided:		
Minimum detectable difference: 10	539	54
Minimum detectable difference: 15	239	24
Confidence level 95%		
Confidence interval approach:		
Two-sided width of interval: 10	680	68
Two-sided width of interval: 15	303	31
Hypothesis testing approach – one-sided:		
Minimum detectable difference: 10	539	54
Minimum detectable difference: 15	239	24
Hypothesis testing approach – two-sided:		
Minimum detectable difference: 10	689	69
Minimum detectable difference: 15	306	31

Source: Calculated by RTI International.

B.7 Sampling and Weights

Generally, for sampling schools, probability proportional to size (PPS) sampling is the most frequently used and recommended sampling technique. With this technique, schools are selected for stage 1 sampling, where the probability of selection of each school is proportional to the number of students in the school (or grade) divided by the number of schools in the desired region or country. The probability of selection of students in the school is the second stage of sampling, where the probability of selection of each student is the number of students to be selected divided by the number of students in the school or grade. Thus, the overall probability of selection of students is the product of these two selection probabilities. While the individual stages present different probabilities of selection of the sampling units for that stage,

the products of the two stages present equal overall probabilities of selection when the number of students selected in each school is the same.

Overall probability of selection is equal to

Stage 1 probability × Stage 2 probability

which is

$$\frac{\#\ of\ pupils\ in\ school\ \times \#\ of\ schools\ selected}{number\ of\ pupils\ in\ region/country} \times \frac{number\ of\ pupils\ selected\ in\ school}{number\ of\ pupils\ in\ school}$$

which can be simplified by:

$$\frac{\cancel{\#\ of\ pupils\ in\ school}\ \times \#\ of\ schools\ selected}{number\ of\ pupils\ in\ region/country} \times \frac{number\ of\ pupils\ selected\ in\ school}{\cancel{number\ of\ pupils\ in\ school}}$$

giving,

$$\frac{\#\ of\ schools\ selected\ \times number\ of\ pupils\ selected\ in\ school}{number\ of\ pupils\ in\ region/country}$$

The final weights are the inverse of the overall probabilities of selection. If the weights are equal, it follows that the weighted mean is the same as the unweighted mean. While in theory, this is what should happen, it usually does not due to fewer pupils sampled in some schools, stratification of schools, replacement schools and so on. However, PPS sampling gives weights that are close to each other, thus reducing sample bias.

ANNEX C: COMPLEX AND CLUSTER SAMPLING

For large education surveys, it is neither cost effective nor practical to randomly select students from the entire target population (i.e., to use simple random sampling). To do so would require a current list of every student in the population of interest, and most ministries of education do not have this information. Even if they did, it would not be cost effective for assessment teams to travel to a particular school to assess just one or two students. Rather, it is more practical and cost effective to randomly sample schools and then sample a cluster of students within each of the selected schools. This sampling methodology is called a *school-based complex sample.* Most EGRA samples are school-based. School-based samples typically involve sampling schools as a whole, and then students (or sometimes first sampling geographical areas, such as districts, and then schools, followed by students). Regardless of these stages of sampling, there is some form of student clustering within the sampled schools.

Many EGRA studies involve *statistical inference.* That is, a random sample of students is drawn from an explicit population of interest and the sampled students' results are used as estimates to infer the results to the population. Called *parametric inferential statistics,* this technique typically incorporates two types of estimates: (1) *point estimates,* or single values calculated from the data to represent unknown parameters; and (2) *precision estimates,* or the range of likely values. (Point and precision estimates are defined more completely below.)

Based on these two estimates and the degrees of freedom, a 95% confidence interval can be calculated and formal statistical analysis can proceed. Note that both types of estimates are directly affected by how the sample is drawn.

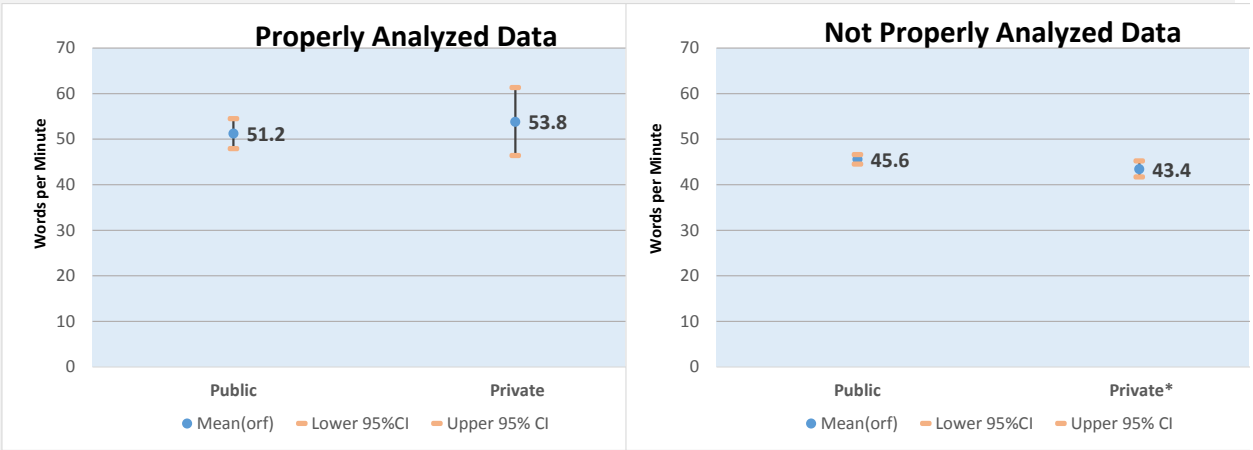
If the sample methodology is not taken into account, the statistical software will assume that the students were chosen by simple random sampling. This will cause all the point estimates of the population parameters to be biased. It will also make all the precision estimates inappropriately low. Combined, this can lead the researcher to incorrectly conclude that there are statistically significant differences among subgroups when in fact there are none, as indicated in **Exhibit C-1.**

Exhibit C-1. Properly analyzed data vs. not properly analyzed data

Oral reading fluency (ORF) was analyzed by school type. The left graph displays data that were analyzed using the proper sample setup. In the graph on the right, the data were analyzed without the correct sample setup. That is, the data were not weighted, and the researcher allowed the statistical analysis software to default to student sampling using simple random sampling.

The mean ORF estimates (words per minute) in the improper analysis are biased to the population, because they assume that only 25% of the student population was from the Java–Bali region of Indonesia, when in fact 55% of the student population was from Java–Bali. Therefore, the sample underrepresents the student population in Java–Bali, causing the mean estimates to be biased.

Furthermore, because the analysis program was allowed to assume simple random sampling, it did not take into account any of the design effects that come with complex samples. This caused the standard error estimate to be extremely low, leading to overconfidence in the biased mean estimate. This researcher might have incorrectly concluded that there was a statically significant difference in students' reading ability in public schools when compared to private schools—whereas the properly analyzed data show that this was not the case.



Source: Grade 2 Indonesia National EGRA (2014). * $p < 0.05$

ANNEX D: SAMPLING FOR IMPACT EVALUATIONS

This annex describes how to use statistical power calculations to appropriately determine the sample necessary to estimate impacts.

Moving from point estimates to estimation of impacts has a large effect on sample size considerations. Statistical power calculation demonstrates how well an evaluation can distinguish real impacts from chance differences. This helps answer two interrelated questions:

1. **Given a sample size, how large an effect would there need to be in order for the study to have a reasonable chance of observing a statistically significant difference between the treatment and control groups?** This question is relevant in a situation where the sample available is set and cannot be altered. For example, a study may be limited to 200 schools in total, and statistical power calculations would allow a researcher to determine how large a change would need to be observed in order to be confident that the change is statistically significant (that is, real). In this instance, the researcher starts with the sample available and then calculates how large the effect size would need to be to detect it confidently.
2. **Given an effect size, how much sample would be needed to ensure with a high probability that the study would detect that effect size, should it result from the intervention?** Here, if there is flexibility in the sample, statistical power calculations can help determine how much sample is necessary to observe a statistically significant change in an observed outcome of a certain magnitude. For example, researchers, USAID, or implementing partners may know, either from the theory of change or from studies of similar interventions, that they expect a certain change in an outcome as a result of the intervention. In this instance, the researcher would start with that expected effect and then determine what sample would be necessary to detect it with confidence.

To determine the appropriate sample sizes for an evaluation, evaluators typically calculate *minimum detectable impacts* (MDIs), which represent the smallest true program impacts—average treatment and control group differences—that can lead to a statistically significant estimate with high probability given a sample size. It is common to standardize MDIs into effect-size units—that is, as percentages of the

standard deviations of the outcome measures. In this instance, the standardized MDI is called the *minimum detectable effect size* (MDES). Scaling impact estimates into standard deviation units facilitates the comparison of findings across outcomes that are measured on different scales.

Mathematically, the MDI formula can be expressed as follows:

$$MDI = Factor * SE(impact),$$

where SE(impact) is the standard error of the impact estimate and Factor is a constant that is a function of the significance and statistical power levels. Factor becomes larger as the significance level is decreased and as the power level is increased. Thus, the MDI rises when we seek to reduce the chances of making Type I and Type II errors. SE(impact) varies according to the impact evaluation design. Generally, larger samples reduce SE(impact) and, thereby, the MDI, making the evaluation "more powerful." Greater power is desirable because the evaluation is more likely to detect substantively meaningful impacts, although greater power typically comes at greater cost.

The formula for the MDES divides the MDI by SD(outcome), the standard deviation of the outcome measure:

$$MDES = MDI / SD(outcome)$$

An MDES is a function of the standard error of the impact estimate, the assumed significance level, and the assumed power level. The significance level is the probability of making a "Type I" error; such an error is a false positive—incorrectly concluding that there is an impact when there is none. A conventional significance level is 5 percent. The power level is one minus the probability of a "Type II" error; such an error is a false negative—failing to detect an impact that truly exists. Evaluations often try to achieve 80 percent power. The goal is to have a small MDES, so that if the study produces an effect larger than the MDES, we call it statistically significant and believe it to be true. All studies include such a calculation, with the formula thoroughly documented to guide decisions around sample size and composition.

Other factors. Many factors that one must consider in developing samples for point estimation are also present in statistical power calculations for impact evaluations. One highly relevant factor for researchers attempting to estimate impacts using EGRA is whether the sample is clustered. Many studies that employ EGRA are clustered, where either schools or communities are first selected to treatment and control and then a certain number of individuals within each cluster are tested. In this

situation, one assumes that individuals within the cluster share certain similarities. For example, children in a classroom, independent of any intervention, may all be taught by the same teacher. This reduces individual variation within the cluster and, in turn, each individual's contribution to the impact estimate. This measure of the degree to which outcomes of individuals within groups are correlated is called the *intraclass correlation coefficient* (ICC). In a situation in which the ICC was closer to 1, adding more individuals would have limited or no effect on the MDES. Rather, the researcher would need to add additional clusters to lower the MDES. In fact, in a clustered study with a high ICC, adding individuals in each cluster will have a minimal positive effect at best. Other factors that impact statistical power calculations are the number of contrasts (or treatment arms or study groups) and whether the test is one-tailed or two-tailed (a one-tailed test would seek to estimate the impact in only one direction).

Finally, it is important to remember that subgroup analysis will have an effect on sample size as well. For example, it may be relevant to understand how a particular intervention impacts boys and girls separately. In this instance, power calculations are done at the subgroup level. Note that in this case, each subgroup is sufficiently large to detect impacts between members of that group alone in the treatment and control groups. In other words, the more subgroups to be analyzed (i.e., disaggregations requested by USAID or other stakeholders), the larger the sample required to do so.

Summary. To summarize, a researcher would use the following principles of statistical power calculations to inform sample size.

1. The larger the sample size, the higher the power (for a clustered study, the larger the number of clusters, the higher the power).
2. Power is higher when the standard deviation of the outcome is small than when it is large.
3. The larger the effect size, the more likely it is that an evaluation would find a significant effect.
4. There is a trade-off between the significance level and power: the more stringent (lower) the significance level, the lower the power.
5. Power is higher with a one-tailed test than with a two-tailed test as long as the hypothesized direction is correct.

ANNEX E: EVALUATING THE TECHNICAL QUALITY OF THE EGRA INSTRUMENT

It is important to evaluate the technical quality of any instrument used to measure student achievement. The EGRA instrument is no exception. The procedures used to conduct these checks come from the field of psychometrics. Traditionally, these procedures have focused on two key concepts: reliability and validity. Teams directing an EGRA application must include a person familiar with psychometrics who can run the necessary checks. Explanations of reliability and validity appear in Section 9.1.2. Below are additional types of analyses that can be considered to establish reliability and validity of instruments.

E.1 Reliability Tests

Test–retest method is an alternative measure to test reliability**.** This approach, which can be conducted as part of the piloting of the EGRA instrument, involves administering the EGRA instrument to the same group of students at two different times (e.g., a week or so apart). The students selected are representative of the target population in key areas such as gender and age, socioeconomic status/home background, cognitive abilities, and so on. The reliability coefficient for test–retest represents the correlation between students' scores on the two administrations of the test. These correlations are ideally also established over the summary measures of the subtasks (percent correct, fluency, etc.), because if the correlations are calculated for individual items within subtasks, the same upward bias introduced by EGRA being timed will be present in test–retest as in Cronbach's alpha.

Upward bias can be illustrated using a numerical example. Suppose one has results for two children—or for the same child after some period of time—as presented in **Exhibit E-1,** where 0s and 1s show "incorrect" or "incorrect," respectively. In neither case was the child able to proceed past word 5. If an analyst were to calculate the correlation for words 1 through 5, the correlation would be very low (0.17). If the analyst were to consider words past 5 as incorrect, and then calculate the correlation for all 10 words, the correlation then would appear as 0.38—much higher, but derived inappropriately, because all the 0s past word 5 would have boosted the correlation artificially.

Exhibit E-1. Sample subtask results for calculating upward bias

Word	Child 1	Child 2 (or Child 1 later)
1	0	0
2	1	1
3	0	1
4	1	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0

Two concerns to be aware of when using the test–retest approach for educational testing:

- First, if the time points are too far apart, there is a likelihood that students will have learned a significant amount of information and that the lack of reliability will actually be a measure of improved student performance.
- Second, limiting the amount of time between assessment administrations can reduce the impact of learning but increase the susceptibility to carryover effects. That is, scores on the second administration are impacted by the fact that the test was recently administered to the same students.

Another test reliability measure is **parallel-forms reliability**. This approach uses two similar forms of the EGRA instrument. In this case, the procedure is to administer Form 1 of the test to each student, followed by the administration of Form 2 to the same students. It is recommended that the order of administration of the forms be reversed for half the selected group. The correlation between the two sets of scores offers a measure of the degree of reliability of EGRA scores across the test forms. This approach is most valuable when multiple forms of assessments are created to measure scores at multiple time points (such as baseline, midterm, and endline). It is important to remember, however, that a strong correlation between two test forms does not ensure that the forms are equivalent or that equating will not be necessary. The same caution about applying the correlations to anything but the summary measures of the subtasks applies here.

E.2 Validity Tests

Criterion-related evidence pertains to the strength of the relationship (correlation) between scores on the EGRA test and those on other measures external to the test. In general, this will involve looking at the relationship between EGRA scores and those on measures of some criteria that the test is expected to predict (e.g., reading comprehension scores in later grades), as well as relationships to other tests hypothesized to measure the same or related constructs (e.g., student scores on other early reading skills tests). Data on these other measures may be collected at the same time as the EGRA data, or they may be collected at a later point in time (but they are collected from the same students). This type of validity evidence will be hard to collect in countries with few standardized measures of student learning outcomes. However, it is worth keeping in mind that extensive research in other countries has demonstrated that EGRA-type instruments show strong relationships (0.7 and above) to the types of external measures provided as examples in this paragraph.

Some test developers recommend that an additional type of evidence be collected as part of test validation: **evidence of the consequences of test score use** on test takers and other stakeholders. This action involves collecting data to determine whether the desired beneficial effects of the test are being realized (e.g., in the case of EGRA, desired benefits include providing policy makers with system-level results on early-reading skill levels so that they can more effectively target resources and training). It also involves collecting evidence of any unintended negative consequences of test score use (e.g., punishing schools that perform poorly on EGRA by withholding resources from them) and taking steps to prevent these adverse outcomes from reoccurring.

ANNEX F: RECOMMENDATIONS AND CONSIDERATIONS FOR CROSS-LANGUAGE COMPARISONS

F.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.

F.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.

F.1.2 Alphasyllabic Languages

Within alphasyllabic (e.g., Hindi, Thai, Sinhala, Lao):

1. Know that the number of vowel or consonant components (phonemic diacritics) within each syllable cluster (*akshara*) affects ability to read (Nag & Perfetti, 2014).
2. Know that the type of phonemic diacritic will affect ability to read (Nag, 2014).
3. Know that nonlinearity of the phonemic constituents within a syllable cluster will affect ease of reading ability.
4. Know that because of the large orthographic set that has to be acquired, the number of years of instruction required to reach the same level of fluency in South and Southeast Asian alphasyllabaries is around five (compared to around three in English) (Nag, 2007).

F.1.3 Non-Roman Alphabetic Languages

Within non-Roman alphabetic languages (e.g., Arabic, Hebrew):

1. Know whether the orthographic depth of the language in question is shallow (e.g., voweled Arabic and Hebrew words) or deep (e.g., unvoweled Arabic and Hebrew words).
2. Know that Arabic is a clear case of diglossia (Ferguson, 1959).
   - Diglossia is the term used to describe a situation in which two varieties of a language are used for socially distinct functions. The sociolinguistic functional distinctness and the subsequent linguistic (phonological, syntactic, morpho-syntactic, and lexical) distance between the two forms of Arabic are believed to impede, or at least to slow, initial acquisition of reading (Abu-Rabia, 2000; Ayari, 1996).
   - The diglossic nature of Arabic is closely related to orthographic depth and to reading fluency.
3. Know that vowels are perceived as naturally attached to the consonants.
4. Know that research on the reading of unvoweled Arabic and Hebrew scripts has shown that reading comprehension in these languages is not related to reading accuracy (Saiegh-Haddad, 2003).
5. Know that letter shapes matter in some orthographies (e.g., Arabic) as children do not see many letter shapes separately.

F.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.

F.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.

F.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.

F.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 F

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 G: COMPARISON OF DATA COLLECTION SOFTWARE

Features	Tangerine	Magpi	SurveyToGo	doForms	Droid Survey	ODK	Command Mobile
Price		Free license for up to 6,000 completed forms/interviews per year. For higher volume, cost varies.	$0.10 to $0.15 per completed form, depending on volume. Additional fee may apply for transmission/storage of high-resolution photos/videos/audio.	$9.95 per month/$99.95 per year per device for professional version. $14.95 per month/$149.95 per year per device for dispatch version.	$60 for one month, $280 for six months, $400 for one year. Unlimited number of devices and 3,000 results uploads per month.	Free	Standard version: $24.99 per month, $64.99 per quarter, $239.99 per year. Advanced version: $69.99 per month.
Compatibility							
Android	yes	yes	yes	yes	yes	yes	yes
iOS	no	yes	no	yes	Yes, using a separate app called iSurvey	Yes, third-party support	yes
Windows Mobile	no	no	yes	Available soon	no	no	yes
Symbian	no	yes	no	no	no	no	no
Blackberry	no	no	no	no	no	no	no
SMS	no	yes	no	no	no	no	no
Instrument development							
Form-based (no programmer expertise needed)	yes	yes	yes	yes	Yes	yes	yes
Offline instrument editing	yes	Yes, it is possible to edit instrument offline in Excel or XForms format and then upload it	no	yes	no	yes	no
Unicode compatible; wide language/script compatibility	yes	yes	yes	yes	yes	yes	yes
User interface for language localization	yes	Choice of five different languages	Yes	English, Spanish, Russian	Yes, choice of about 15 languages	yes	Does not say
EGRA core subtask templates	yes	no	no	no	no	no	no
EGMA core subtask templates	yes	no	no	no	no	no	no
Possible to create EGRA instrument?	yes	No, not without contracting customization services	Yes, but not without some training	Yes, with any of the paid versions	Yes (but it would not be easy as this grid table is intended for several rows of questions with the same multiple answer choices)—i.e., it's not possible to actually label the items in the grid	Yes, has been done before	Demo not available online
Print forms?	yes	no	yes	no	no	yes	no

Features	Tangerine	Magpi	SurveyToGo	doForms	Droid Survey	ODK	Command Mobile
Data collection							
Offline data collection	yes	yes	yes	yes	yes	yes	yes
Text/numerical data	yes	yes	yes	yes	yes	yes	yes
Timed survey data	yes	no	yes	yes	no	yes	no
Grid tables	yes	no	yes	yes	yes	yes	yes
Single/multiple-choice answers	yes	yes	yes	yes	yes	yes	yes
GPS location	yes	yes	yes	yes	yes	yes	yes
Screenshots	yes	yes	yes	yes	yes	yes	yes
Camera	no	no	yes	yes	yes	yes	yes
Video	no	no	yes	yes	no	yes	yes
Audio	no	no	yes	yes	no	yes	yes
Barcode	no	no	yes	yes	yes	yes	yes
Signature	no	no	yes	yes	yes	yes	yes
Logic features							
Skip logic	yes	yes	yes	yes	yes	yes	yes
Custom validation	yes	yes	yes	yes	no	yes	yes
Conditional form display	yes	yes	yes	yes	no	yes	yes
Looping	no	yes	yes	yes	no	yes	no
Question branching	yes	yes	yes	yes	no	yes	no
Uploading data							
WiFi	yes	yes	yes	yes	yes	yes	yes
Cellular	yes	yes	yes	yes	yes	yes	yes
Cable	no	yes	yes	yes	yes	yes	yes
Device-to-device backup	yes	Yes, via memory card	no	no	no	Yes, with third-party support	no
Export data to							
Excel	yes	yes	yes	yes	yes	yes	yes
RSS	no	no	no	no	no	no	yes
SPSS	yes	no	yes	no	yes	no	no
MS Word	no	yes	yes	yes	no	no	no
MS Access	no	yes	yes	yes	no	no	no
XML	no	yes	yes	yes	no	yes	no
HTML	no	no	no	yes	no	no	no
PDF	no	yes	no	yes	no	no	no
GoogleDocs	no	no	no	yes	no	yes	no
OpenOffice	no	no	no	yes	no	no	no
Data storage							
Amount of storage included		Unlimited cloud storage	20 MB for attachments; more can be purchased. Unlimited cloud storage	Unlimited cloud storage	Unlimited cloud storage	Unlimited	250GB included
Encryption during transfer?	no	Encrypted to 256-bit strong standard	Yes, at extra cost	Yes—SSL encryption	Yes—SSL encryption	yes	yes
Open source?	yes	No, but API available to Enterprise customers	no	API available but not open source	Results API is available on request	yes	no
Kiosk mode?	no	yes	yes	no	yes	Yes, with third-party support	no

ANNEX H: COMPARISON OF PAPER VS. ELECTRONIC EGRA INSTRUCTIONS

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
▪ **General Instructions**		
▪ Establish a playful and relaxed rapport with the child through a short conversation (see example topics below). The child should perceive the assessment almost as a game to be enjoyed rather than a test. Use this time to identify in what language the child is most comfortable communicating. Read aloud slowly and clearly ONLY the sections in boxes.	Establish a playful and relaxed rapport with the child through a short conversation (see example topics below). The child should perceive the assessment almost as a game to be enjoyed rather than a test. Use this time to identify in what language the child is most comfortable communicating. Read aloud slowly and clearly ONLY the sections in boxes.	**Good morning. My name is ____ and I live in _____. I'd like to tell you a little bit about myself.** [Number and ages of children; favourite sport, radio or television program, etc.] **1. What do you like to do when you are not in school?** [Wait for response; if student is reluctant, ask question 2, but if they seem comfortable continue to oral assent]. **2. What games do you like to play?**
Oral assent		
If oral asset is not obtained, thank the child and move on to the next child, using this same form.	If oral assent is not obtained, thank the child, press "no," then press "save" and "perform another assessment."	• **Let me tell you why I am here today. I work with the Ministry of Education and we are trying to understand how children learn to read. You were picked by chance.** • **We would like your help in this. But you do not have to take part if you do not want to.** • **We are going to play a reading game. I am going to ask you to read letters, words and a short story out loud.** • **Using this stopwatch/device/gadget, I will see how long it takes you to read.** • **This is NOT a test and it will not affect your grade at school.**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
		• **I will also ask you other questions about your family, like what language your family uses at home and some of the things your family has.** • **I will NOT write down your name so no one will know these are your answers.** • **Once again, you do not have to participate if you do not wish to. Once we begin, if you would rather not answer a question, that's all right.** **Do you have any questions? Are you ready to get started?**
Orientation to Print		
Show the child the paragraph segment on the last page of the student assessment. Read the instructions to the child. Record the child's response before moving to the next instruction.	Show the child the paragraph segment on the last page of the student assessment. Read the instructions to the child. Record the child's response before moving to the next instruction.	**I don't want you to read this now.** **On this page, where would you begin to read? Show me with your finger.** **Now show me in which direction you would read next.** **When you get to the end of the line, where would you read next?**
Letter Sound Identification		
Show the child the sheet of letters in the student stimuli booklet as you read the instructions below to the child. **[INSERT INSTRUCTIONS IN RIGHT-HAND COLUMN TO BE READ TO CHILD]** Start the timer when the child reads the first letter. • Follow along with your pencil and clearly mark any incorrect letters with a slash (/).	Show the child the sheet of letters in the student stimuli booklet as you read the instructions. Start the timer when the child reads the first letter. Follow along on your screen and mark any incorrect letters by touching that letter on the screen—it will turn blue. Mark self-corrections as correct by touching the letter again—it will return to gray.	**Here is a page full of letters of the English alphabet. Please tell me the SOUNDS of as many letters of the alphabet as you can. Not their names, but their sounds.** **For example, the sound of this letter** [point to the letter T] **is /t/.** **Let's practice: Tell me the sound of this letter** [point to the letter M]**:** [If the child responds correctly, say:] **Good, the sound of this letter is /m/.** [If the child does not respond correctly, say:] **The sound of this letter is /m/.**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
• Count self-corrections as correct. If you already marked the self-corrected word as incorrect, circle it (ø) and continue. • If the pupil skips an entire line, draw a line through it on the protocol. • Stay quiet, except if the child hesitates for 3 seconds. Point to the next letter and say, "Please go on." Mark the skipped letter as incorrect. • If the student provides the letter name rather than the sound, pronounce the letter sound and say: "Please tell me the SOUND of the letter." This prompt may be given only once during the subtask. When the timer reaches 0, say "stop." Mark the final word read with a bracket ( ]). Early stop rule: 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. [AT THE BOTTOM OF THE PAGE, INCLUDE THE FOLLOWING LINES] Time remaining on stopwatch at completion (number of SECONDS) ☐ Exercise discontinued because the child had no correct answers in the first line ☐	Stay quiet, except if the child hesitates for 3 seconds. Then point to the next letter and say, "Please go on." Mark the skipped letter as incorrect. If the student provides the letter name rather than the sound, say: "Please tell me the SOUND of the letter." This prompt may be given only once during the exercise. If the timer runs out before the last item is read, the screen will flash red and the timer will stop. Tell the child to stop. Mark the final letter read by touching it so that a red bracket appears. Then press "Next." If the child reaches the last item before the screen flashes red, stop the timer as soon as the child reads the last letter. Touch the last letter so the red bracket appears. Then press "Next." Early stop rule: If the child does not provide a single correct response for the first 10 items, the screen will flash red and the timer will stop. Say, "Thank you!", discontinue this subtask, and go on to the next subtask.	**Now try another one: Tell me the sound of this letter** [point to the letter S]**:** [If the child responds correctly say:] **Good, the sound of this letter is /s/.** [If the child does not respond correctly, say:] **The sound of this letter is /s/.** **When I say "Begin," start here** [point to first letter] **and go across the page** [point]. **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.** [AFTER SUBTASK ITEMS] **Good effort! Let's go on to the next section.**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
Letter Name Identification		
Show the child the sheet of letters in the student stimuli booklet as you read the instructions below to the child. **[INSERT INSTRUCTIONS IN RIGHT-HAND COLUMN TO BE READ TO CHILD]** Start the timer when the child reads the first letter. • Follow along with your pencil and clearly mark any incorrect letters with a slash (/). • Count self-corrections as correct. If you already marked the self-corrected word as incorrect, circle it (ø) and continue. • If the pupil skips an entire line, draw a line through it on the protocol. • Stay quiet, except if the child hesitates for 3 seconds. Point to the next letter and say, "Please go on." Mark the skipped letter as incorrect. • If the student gives you the letter SOUND, rather than the name, provide the letter name and say: "Please tell me the NAME of the letter." This prompt may be given only once during the subtask. When the timer reaches 0, say "stop." Mark the final word read with a bracket ( ]). Early stop rule: If the child does not provide a single correct response on the first line (10 items), say "Thank you!", discontinue this subtask, check	Show the child the sheet of letters in the student stimuli booklet as you read the instructions. Start the timer when the child reads the first letter. Follow along on your screen and mark any incorrect letters by touching that letter on the screen—it will turn blue. Mark self-corrections as correct by touching the letter again—it will return to gray. Stay quiet, except if the child hesitates for 3 seconds. Then point to the next letter and say, "Please go on." Mark the skipped letter as incorrect. If the student provides the letter sound rather than the name, say: "Please tell me the NAME of the letter." This prompt may be given only once during the exercise. If the timer runs out before the last item is read, the screen will flash red and the timer will stop. Tell the child to stop. Mark the final letter read by touching it so that a red bracket appears. Then press "Next." If the child reaches the last item before the screen flashes red, stop the timer as soon as the child reads the last letter. Touch the last letter so the red bracket appears. Then press "Next."	**Here is a page full of letters of the ENGLISH alphabet. Please tell me the NAMES of as many letters of the alphabet as you can. Not their sounds, but their names.** **For example, the name of this letter** [point to the letter T] **is "T".** **Let's practice: Tell me the name of this letter** [point to the letter M]**:** [If the child responds correctly say]: **Good, the name of this letter is "em."** [If the child does not respond correctly, say]: **The name of this letter is "em".** **Now try another one: Tell me the name of this letter** [point to the letter S]**:** [If the child responds correctly, say]: **Good, the name of this letter is "es."** [If the child does not respond correctly, say]: **The name of this letter is "es."** **When I say "Begin," start here** [point to first letter] **and go across the page** [point]. **Point to each letter and tell me the name 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.** [AFTER SUBTASK ITEMS] **Good effort! Let's go on to the next section.**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
the box at the bottom, and go on to the next subtask. **[AT THE BOTTOM OF THE PAGE, INCLUDE THE FOLLOWING LINES]** Time remaining on stopwatch at completion (number of SECONDS) ☐ Exercise discontinued because the child had no correct answers in the first line ☐	Early stop rule: If the child does not provide a single correct response for the first 10 items, the screen will flash red and the timer will stop. Say, "Thank you!", discontinue this subtask, and go on to the next subtask.	
Phonemic Awareness – Task 1: Initial Sound Identification		
This is NOT a timed exercise and THERE IS NO STUDENT SHEET. Remove the pupil stimuli booklet from the child's view. Read the instructions to the child and conduct the examples. **[INSERT INSTRUCTIONS TO CHILD FROM RIGHT-HAND COLUMN]** Read the prompt and then pronounce the word a second time. Pronounce each word slowly. Accept as correct only the isolated sound. Mark the box that corresponds to the child's answer. If the child does not respond after 3 seconds, mark as "No response" and say the next prompt. Early stop rule: If the child responds incorrectly or does not respond to the first five words, say "Thank you!, discontinue this subtask, check the	This is NOT a timed exercise and THERE IS NO STUDENT SHEET. Read the instructions to the child and conduct the examples. Read the prompt and then pronounce the word a second time. Follow along on your screen and mark each item as either "Correct" or "Incorrect." Your selection will turn yellow. If the child does not respond after 3 seconds, mark as "No response" and say the next prompt. Early stop rule: If the child does not provide a single correct response for the first 5 items, the screen will flash red and the timer will stop. Say, "Thank you!", discontinue this subtask, and go on to the next subtask.	**This is a listening exercise. I want you to tell me the first sound of each word. For example, in the word "pot," the first sound is /p/. I would like you to tell me the first sound you hear in each word. I will say each word <u>two times</u>. Listen to the word, then tell me the very first sound in that word.** **Let's practice. What is the first sound in "mouse"? "mouse"?** [If the child responds correctly, say:] **Very good, the first sound in "mouse" is /m/.** [If the child does not respond correctly, say]: **Listen again: "mouse." The first sound in "mouse" is /mmm/.** **Now let's try another one: What is the first sound in "day"? "day"?** [If the child responds correctly, say]: **Very good, the first sound in "day" is /d/.**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
box at the bottom of the page, and go on to the next subtask. **[INSERT AT BOTTOM OF PAGE]** Exercise discontinued because the child had no correct answers in the first five words ☐		[If the child does not respond correctly, say]: **Listen again: "day". The first sound in "day" is /d/.** **Ready? Let's begin.** [INSERT DIRECTLY ABOVE THE WORD LIST] **What is the first sound in "___"? "____"?** [INSERT DIRECTLY BELOW THE WORD LIST] **Good effort! Let's go on to the next section.**
Phonemic Awareness **Task 2: Phoneme Segmentation**		
This is NOT a timed exercise and THERE IS NO PUPIL STIMULI. Remove the pupil stimuli booklet from the child's view. Read the instructions to the child and conduct the examples. Read aloud each word twice and have the child say the sounds. When you practice, remember to say the "clipped" sounds /p/. DO NOT say "puh." DO NOT say "pay." **[INSERT DIRECTIONS TO CHILD FROM RIGHT-HAND COLUMN]** Pronounce each word slowly. Do not break the word into individual sounds. Say each word only twice. If the child provides a letter name rather than a sound, say: "Please tell me the sounds in the	This is NOT a timed subtask and there is not a pupil stimulus. Remove the pupil stimuli booklet from the child's view. Read the instructions to the child and conduct the examples. When you practice, remember to say the "clipped" sounds /p/. DO NOT say "puh." DO NOT say "pay." Pronounce each word slowly. Do not break the word into individual sounds. Say each word only twice. Touch the screen only for each sound that is incorrect—It will turn yellow. If the child provides a letter name rather than a sound, say: "Please tell me the sounds in the words, not the letters." This prompt may be given only once during the subtask.	**This is a listening activity. You know that each letter has a sound. For example, "pot," can be sounded as /p/ /o/ /t/. I will say English words twice. Listen to the word then tell me all the sounds in the word.** **Let's practice. What are the sounds in "fan" - "fan"?** [If the child responds correctly, say]: **Very good! The sounds in "fan" are /f/ /a/ /n/.** [If the child does not respond correctly, say]: **The sounds in "fan" are: /f/ /a/ /n/. Now it's your turn. Tell me the sounds in "fan."** [Wait 3 seconds for the child to respond.] **Let's try another one. What are the sounds in "miss" - "miss"?** [If the child responds correctly, say]: **Very good! The sounds in "miss" are /m/ /i/ /s/.**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
words, not the letters." This prompt may be given only once during the subtask. Put a slash (/) through each incorrect phoneme, well as any phonemes that the child does not say. If the child does not respond after 3 seconds, mark all sounds as incorrect and proceed to the next word. If the child provides all the correct sounds, tick the box "All correct." Early stop rule: If the child does not provide a single correct response for the first 5 items, say "Thank you!, discontinue this subtask, check the box at the bottom of the page, and go on to the next subtask. **[INSERT AT BOTTOM OF PAGE]** Exercise discontinued because the child had no correct answers in the first five words ☐	If the child does not respond after 3 seconds, mark all sounds as incorrect and proceed to the next word. If the child provides all the correct sounds, tick the box "All correct." Early stop rule: If the child does not provide a single correct response for the first 5 items, the screen will flash red and the timer will stop. Say, "Thank you!", discontinue this subtask, and go on to the next subtask.	[If the child does not respond correctly, say]: **The sounds in "miss" are: /m/ /i/ /s/. Now it's your turn. Tell me the sounds in "miss."** [Wait 3 seconds for the child to respond.] **Okay. Let's start. I will say a word twice. Listen to the word, and then tell me the sounds in that word. Ready? Let's begin.** **What are the sounds in "_____"? "_____"?** [Say each word twice.]
Syllable Identification		
Show the child the sheet of syllables in the student stimuli booklet. Read the instructions to the child and conduct the examples. [INSERT INSTRUCTIONS TO READ TO THE CHILD FROM THE RIGHT-HAND COLUMN] Start the timer when the child reads the first syllable.	Read the instructions. Be sure to show the child the sheet of syllables, pointing to the first syllable when you say "Begin." Start the timer when the child reads the first syllable. Follow along on your screen and mark any incorrect syllable by touching that syllable on the screen—it will turn blue. Mark self-	**Here are some syllables. I would like you to read as many syllables as you can. Do not spell the syllables, but read them. For example, this syllable is: "ba."** **Let's practice: Please read this syllable** [point to the syllable "**lo**"]**:** [If the child responds correctly, say:] **Good, this syllable is "lo."** [If the child does not respond correctly, say:] **This syllable is "lo."**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
• Follow along with your pencil and clearly mark any incorrect syllables with a slash (/). • Count self-corrections as correct. If you already marked the self-corrected syllable as incorrect, circle it (ø) and continue. • If the pupil skips an entire line, draw a line through it on the protocol. • Stay quiet, except if the child hesitates for 3 seconds. Then point to the next letter and say, "Please go on." Mark the skipped syllable as incorrect. When the timer reaches 0, say "stop." Mark the final word read with a bracket ( ]). Early stop rule: 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. [AT THE BOTTOM OF THE PAGE, INCLUDE THE FOLLOWING LINES] Time remaining on stopwatch at completion (number of SECONDS) ☐ Exercise discontinued because the child had no correct answers in the first line ☐	corrections as correct by touching the syllable again—it will return to gray. Stay quiet, except if the child hesitates for 3 seconds. Then point to the next syllable and say, "Please go on." Mark the skipped syllable as incorrect. If the timer runs out before the last item is read, the screen will flash red and the timer will stop. Tell the child to stop. Mark the final syllable read by touching it so that a red bracket appears. Then press "Next." If the child reaches the last item before the screen flashes red, stop the timer as soon as the child reads the last syllable. Touch the last syllable so the red bracket appears. Then press "Next." Early stop rule: If the child does not provide a single correct response for the first 10 items, the screen will flash red and the timer will stop. Say, "Thank you!", discontinue this subtask, and go on to the next subtask.	**Now try another one: Please read this syllable** [point to the syllable "mu"]**:** [If the child responds correctly say:] **Good, this syllable is "mu."** [If the child does not respond correctly, say:] **This syllable is "mu."** **When I say "Begin," start here** [point to first syllable] **and go across the page** [point]. **Point to each syllable and read it in a loud voice. Read as quickly and carefully as you can. If you come to a syllable you do not know, go on to the next syllable. Put your finger on the first syllable. Ready? Begin.** [INSERT DIRECTLY AFTER THE SYLLABLE LIST] **Good effort! Let's go on to the next section.**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
Syllable Segmentation		
This is NOT a timed subtask and THERE IS NO PUPIL STIMULUS. Remove the pupil stimuli booklet from the child's view. Read the instructions to the child and conduct the examples. Read aloud each word twice and have the child say the syllables. Remember, when you practice, say the "clipped" syllables, such as "ba." **[INSERT DIRECTIONS TO CHILD FROM RIGHT-HAND COLUMN]** Pronounce each word slowly. Do not break the word into individual syllables. Say each word only twice. If the child provides the word rather than the syllable, say, "Please tell me all the syllables in the word." This prompt may be given only once during the subtask. Put a slash (/) through each incorrect syllable as well as any syllables that the child does not say. If the child does not respond after 3 seconds, mark all the syllables as incorrect and proceed to the next word. If the child provides all the correct sounds, tick the box "All correct." Early stop rule: If the child does not provide a single correct response for the first 5 items, say	This is NOT a timed subtask and there is not a pupil stimulus. Remove the pupil stimuli booklet from the child's view. Read the instructions to the child and conduct the examples. When you practice, remember to say the "clipped" syllables, such as "ba." Pronounce each word slowly. Do not break the word into individual syllables. Say each word only twice. Touch the screen only for each syllable that is incorrect—It will turn yellow. If the child provides the word rather than the syllable, say, "Please tell me all the syllables in the word." This prompt may be given only once during the subtask. If the child does not respond after 3 seconds, mark all the syllables as incorrect and proceed to the next word. If the child provides all the correct sounds, tick the box "All correct." Early stop rule: If the child does not provide a single correct response for the first 5 items, the screen will flash red and the timer will stop. Say, "Thank you!", discontinue this subtask, and go on to the next subtask.	**This is a listening activity. You know that each word has syllables. For example, "yesterday" can be sounded as /yes/ /ter/ /day/. I will say English words twice. Listen to the word, then tell me all the syllables in the word.** **Let's practice. What are the syllables in "rabbit" - "rabbit"?** [If the child responds correctly, say:] **Very good! The sounds in "rabbit" are /rab/ /bit/.** [If the child does not respond correctly, say:] **The sounds in "rabbit" are: /rab/ /bit/. Now it's your turn. Tell me the sounds in "rabbit."** [Wait 3 seconds for the child to respond.] **Let's try another one. What are the syllables in "wonderful" - "wonderful"?** [If the child responds correctly, say]: **Very good! The sounds in "wonderful" are /won/ /der/ /ful/.** [If the child does not respond correctly, say]: **The syllables in "wonderful" are: /won/ /der/ /ful/. Now it's your turn. Tell me the sounds in "wonderful."** [Wait 3 seconds for the child to respond.] **Okay. Let's start. I will say a word twice. Listen to the word then tell me the syllables in that word. Ready? Let's begin.** **What are the syllables in "_____"? "_____"?** [Say each word twice.]

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
"Thank you!, discontinue this subtask, check the box at the bottom of the page, and go on to the next subtask. **[INSERT AT BOTTOM OF PAGE]** Exercise discontinued because the child had no correct answers in the first five words ☐		
Familiar Word Reading		
Show the child the sheet of words in the student stimuli booklet. Read the instructions to the child and conduct the examples. **[INSERT DIRECTIONS TO CHILD FROM RIGHT-HAND COLUMN]** Start the timer when the child reads the first word. • Follow along with your pencil and clearly mark any incorrect words with a slash (/). • Count self-corrections as correct. If you already marked the self-corrected word as incorrect, circle it (ø) and continue. • If the pupil skips an entire line, draw a line through it on the protocol. • Stay quiet, except if the child hesitates for 3 seconds. Point to the next word and say, "Please go on." Mark the skipped word as incorrect. When the timer reaches 0, say "stop." Mark the final word read with a bracket ( ]).	Show the child the sheet of words in the student stimuli booklet. Read the instructions. Start the timer when the child reads the first word. Follow along on your screen and mark any incorrect word by touching that word on the screen—It will turn blue. Mark self-corrections as correct by touching the word again—It will return to gray. Stay quiet, except if the child hesitates for 3 seconds. Then point to the next word and say, "Please go on." Mark the skipped word as incorrect. If the timer runs out before the last item is read, the screen will flash red and the timer will stop. Tell the child to stop. Mark the final word read by touching it so that a red bracket appears, then press "Next." If the child reaches the last item before the screen flashes red, stop the timer as soon as the child reads the last word. Touch the last	**Here are some words in ENGLISH. I would like you to read as many words as you can. Do not spell the words, but read them. For example, this word is: "cat."** **Let's practice: Please read this word** [point to the word **"sick"**]**:** [If the child responds correctly say:] **Good, this word is "sick."** [If the child does not respond correctly say:] **This word is "sick."** **Now try another one: Please read this word** [point to the word **"made"**]**:** [If the child responds correctly say:] **Good, this word is "made."** [If the child does not respond correctly say:] **This word is "made."** **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.** [INSERT AFTER THE WORD LIST]

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
Early stop rule: 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. [AT THE BOTTOM OF THE PAGE, INCLUDE THE FOLLOWING LINES] Time remaining on stopwatch at completion: ☐ Exercise discontinued because the child had no correct answers in the first line ☐	word so the red bracket appears, then press "Next." Early stop rule: If the child does not provide a single correct response for the first 5 words, the screen will flash red and stop the timer. Say, "Thank you!", discontinue this subtask, and go on to the next subtask.	**Good effort! Let's go on to the next section.**
Nonword Reading		
Show the child the sheet of words in the student stimuli booklet. Read the instructions to the child and conduct the examples. **[INSERT DIRECTIONS TO CHILD FROM RIGHT-HAND COLUMN]** Start the timer when the child reads the first nonword. • Follow along with your pencil and clearly mark any incorrect words with a slash (/). • Count self-corrections as correct. If you already marked the self-corrected letter as incorrect, circle it (ø) and continue. • If the pupil skips an entire line, draw a line through it on the protocol. • Stay quiet, except if the child hesitates for 3 seconds. Then point to the next letter and say, "Please go on." Mark the skipped word as incorrect.	Show the child the sheet of nonwords in the student stimuli booklet. Read the instructions. Start the timer when the child reads the first nonword. Follow along on your screen and mark any incorrect nonwords by touching that word on the screen—It will turn blue. Mark self-corrections as correct by touching the word again—It will return to gray. Stay quiet, except if the child hesitates for 3 seconds. Then point to the next word and say, "Please go on." Mark the skipped word as incorrect. If the timer runs out before the last item is read, the screen will flash red and the timer will stop. Tell the child to stop. Mark the final	**Here are some made-up words in English. I would like you to read as many as you can. Do not spell the words, but read them. For example, this made-up word is: "ut."** **Let's practise: Please read this word** [point to the word: **"dif"**]. [If the child responds correctly:] **Good, this word is "dif."** [If the child does not respond correctly, say:] **This made-up word is "dif."** **Now try another one: please read this word** [point to the next word: **mab**]. [If the child responds correctly, say:] **Good, this made-up word is "mab."** [If the child does not respond correctly say:] **This made-up word is "mab."** **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**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
When the timer reaches 0, say "stop." Mark the final word read with a bracket ( ]). Early stop rule: 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. [AT THE BOTTOM OF THE PAGE, INCLUDE THE FOLLOWING LINES] Time remaining on stopwatch at completion ☐ Exercise discontinued because the child had no correct answers in the first line ☐	nonword read by touching it so that a red bracket appears, then press "Next." If the child reaches the last item before the screen flashes red, stop the timer as soon as the child reads the last nonword. Touch the last nonword so the red bracket appears, then press "Next." Early stop rule: If the child does not provide a single correct response for the first 5 words, the screen will flash red and stop the timer. Say, "Thank you!", discontinue this subtask, and go on to the next subtask.	**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.** [AFTER THE WORD LIST] **Good effort! Let's go on to the next section.**
Oral Reading Fluency / Passage Reading		
Show the child the story in the student stimuli booklet. Read the instructions to the child. **[INSERT DIRECTIONS TO CHILD FROM RIGHT-HAND COLUMN]** Start the timer when the child reads the first word. • Follow along with your pencil and clearly mark any incorrect words with a slash (/). • Count self-corrections as correct. If you already marked the self-corrected word as incorrect, circle it (ø) and continue. • Stay quiet, except if the child hesitates for 3 seconds. Point to the next word and say, "Please go on." Mark the skipped word as incorrect.	Show the child the story in the stimuli booklet. Read the instructions. Read the instructions to the child and conduct the examples. Start the timer when the child reads the first word. Follow along on your screen and mark any incorrect word by touching that word on the screen—It will turn blue. Mark self-corrections as correct by touching the word again—It will return to gray. Stay quiet, except if the child hesitates for 3 seconds. Then point to the next word and say, "Please go on." Mark the skipped word as incorrect.	**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.**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
• If the pupil skips an entire line, draw a line through it on the protocol. When the timer reaches 0, say "stop." Mark the final word read with a bracket ( ]). After the child is finished reading, REMOVE the passage from in front of the child. Early stop rule: If the child does not provide a single correct word on the first line of text, say "Thank you!", discontinue this subtask and check the box at the bottom. Do not ask any comprehension questions.	If the timer runs out before the last item is read, the screen will flash red and the timer will stop. Tell the child to stop. Mark the final word read by touching it so that a red bracket appears, then press "Next." If the child reaches the last item before the screen flashes red, stop the timer as soon as the child reads the last word. Touch the last word so the red bracket appears, then press "Next." Remove the passage from in front of the child. Early stop rule: If the child does not read a single word correctly from among the words necessary to answer the first comprehension question, the screen will flash red and the timer will stop. Say, "Thank you!", discontinue this subtask, and go on to the next subtask.	
Reading Comprehension		
Ask the child only the questions related to the text read. A child must read **all the text** that corresponds with a given question. Mark the child's response and continue to the next question. Mark as correct only responses that are the same or similar in meaning to the answers provided next to each question. If the child does not respond to a question after 10 seconds, continue to the next question. Do not repeat the question.	Before asking the questions, REMOVE the passage from in front of the child**.** Ask the child all the questions presented on the screen. They are automatically aligned with how far the child has read in the oral reading passage. Mark each question as either "Correct" or "Incorrect." Your selection will turn yellow. If the child does not answer after 10 seconds, mark as "No response," and continue to the next question.	**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.**

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
Read the instructions to the child. **[INSERT DIRECTIONS TO CHILD FROM RIGHT-HAND COLUMN]**	Responses with a meaning similar to those provided should be marked correct. If a child says "I don't know," mark as "Incorrect."	
Listening Comprehension		
Remove the pupil stimuli booklet from the child's view. Read the directions to the child. **[INSERT DIRECTIONS TO CHILD FROM RIGHT-HAND COLUMN]** This is NOT a timed subtask. Read the entire passage aloud to the child **ONE TIME ONLY**. Read slowly (about 1 word per second). Ask all of the questions. Do not allow the child to look at the passage or the questions. Mark the child's response and continue to the next question. Mark as correct only responses that are the same or similar in meaning to the answers provided next to each question. If the child does not respond to a question after 10 seconds, continue to the next question. Do not repeat the question.	Remove the pupil stimuli booklet from the child's view. Read the directions to the child. This is NOT a timed subtask. Read the entire passage aloud to the child **ONE TIME ONLY**. Read slowly (about 1 word per second). Ask all of the questions. Do not allow the child to look at the passage or the questions. Mark each question as either "Correct" or "Incorrect." Your selection will turn yellow. If the child does not answer after 10 seconds, mark as "No response," and continue to the next question. Do not repeat the question. Responses with a meaning similar to those provided should be marked correct. If a child says "I don't know," mark as "Incorrect."	**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.** [INSERT AFTER QUESTIONS] **Good effort! Let's go on to the next section.**
Dictation		
Turn the student response form to the last, lined page for writing and place it in front of the student. The student will write the dictation sentence on the lined page of the response form.	Turn the student response form to the last, lined page for writing and place it in front of the student. The student will write the dictation sentence on the lined page of the response form.	**I am going to read you a short sentence. Please listen carefully. I will read the whole sentence once. Then I will read it in parts so you can write what you hear. I will read it a third time so that you can check your work. Do you understand what you are to do?** [INSERT AFTER TEXT]

Instructions to assessor: PAPER	Instructions to assessor: ELECTRONIC	Instructions to child (same for both paper/electronic)
Take the student stimulus sheet and turn to the last page, where you will find the same instructions as below. Hand a pencil to the child. Say to the child: **[INSERT DIRECTIONS TO CHILD FROM RIGHT-HAND COLUMN]** Read the following sentence ONCE, at about one word per second. **[INSERT DICTATION SENTENCE]** Then read the sentence a second time, pausing 10 seconds between groups of words. **[INSERT SAME DICTATION SENTENCE, WITH WORDS SOMEWHAT SPREAD OUT TO INDICATE THE PAUSE BETWEEN THEM.]** After 15 seconds, read the whole sentence again. **[INSERT SAME DICTATION SENTENCE]** Wait up to 15 more seconds for the child to finish writing, then discontinue the subtask.	Take the student stimulus sheet and turn to the last page, where you will find the same instructions as below. Hand a pencil to the child. Read the sentence once, at about one word per second. Then read the sentence a second time, pausing 10 seconds between groups of words. After 15 seconds, read the whole sentence again. Wait up to 15 more seconds for the child to finish writing, then discontinue the subtask.	**Good effort! Let's go on to the next section.**

ANNEX I: SAMPLE ASSESSOR TRAINING AGENDA

Training EGRA Data Collectors

Day &Time	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Daily Objectives:	• Understand purpose of EGRA • Be able to apply administration and scoring rules on paper	• Understand tablet functions and administration • Be able to upload data	• Improve test administration skills • Become familiar with questionnaire administration	• Polish EGRA administration skills and scoring accuracy	• Polish EGRA administration skills and scoring accuracy	• Supervisor training • Team preparations
8:30-9:00 a.m.	• Welcome/introductions	• Review of Day 1	School visit 1: EGRA practice	School visit 2: EGRA + questionnaires	School visit 3: EGRA + questionnaires	• Supervisor training • Team preparations for data collection
9:00-10:30 a.m.	• Overview of EGRA: purpose, instrument content • Purpose of EGRA in this context	• Overview of basic tablet functions				
10:30-11:00 a.m.	*Break*	*Break*				
11:00-1:00 p.m.	• Instrument overview • Demonstration and practice of subtasks	• Practice EGRA on tablets (small groups)				
1:00-2:00 p.m.	*Lunch*					
2:00-3:30 p.m.	• Continued demonstration and practice of subtasks • Pupil questionnaire	• Tablet functionality issues • Uploading data	• School visit debrief • *Additional survey instruments if administered*	• School visit debrief • Discuss Assessor Accuracy Measure 2 results	• School visit debrief • Discuss Assessor Accuracy	

Day &Time	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
				• Practice EGRA on tablets in pairs (key tasks/issues)	Measure 2 results • Data collection logistics	
3:30-3:45 p.m.			*Break*			
3:45-5:30 p.m.	• Continued whole and small-group practice and correction	• EGRA sampling procedures • School visit logistics	• Practice EGRA on tablets in pairs (key tasks/issues) • Assessor Accuracy Measure • Review school visit logistics	• Assessor Accuracy Measure 2 *Additional survey instruments if administered*	• Assessor Accuracy Measure 3	

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 J: 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.

J.1 Data Preparation

Exhibit J-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 J-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													1	1
7	apanjirr1	0	42													1	1
8	ashooirr1	0	42													1	0
9	dmtitirr1	0	40													1	1
10	hseleirr1	0	41													1	1
11	ikiwairr1	0	41													1	1
12	jmasairr1	0	41													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.

J.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 expresses this figure as a percentage of the number of items in the subtask/assessment, as shown in **Exhibit J-2.**

Exhibit J-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
jurasirr1	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
mndolirr1	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 includes the following:

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

J.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.

J.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 J

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 K: SAMPLE PLANS FOR FIELD-BASED INTERRATER RELIABILITY TESTING

This annex accompanies Save the Children's protocol in Section 8.7, which describes how to assess interrater reliability (IRR) on an ongoing basis during an EGRA survey. The charts below show a systematic way to vary the assessment pairs for the first assessment of the day at each school, for teams consisting of three, four, or five assessors. While the total sample size needed for IRR will vary based on the survey design (i.e., the number of schools and students being assessed overall), it is recommended that at minimum, 150 students be double assessed. A sample size of less than 100 for IRR will likely not yield useful information.

All charts and supporting text:

© 2015 by Save the Children. Used by permission. All rights reserved

Inter-rater Reliability Scenarios by Number of Assessors

Schedule with 3 assessors

School	Assessor Pairing	Assesses & Records	Listens & Records
School 1	A & B C	Assessor A Assessor C	Assessor B ----
School 2	B & C A	Assessor B Assessor A	Assessor C ----
School 3	C & A B	Assessor C Assessor B	Assessor A ----
School 4	A & C B	Assessor A Assessor B	Assessor C
School 5	B & A C	Assessor B Assessor C	Assessor A ----
School 6	C & B A	Assessor A Assessor C	Assessor B ----
Etc …			

Schedule with 4 assessors

School	Assessor Pairing	Assesses & Records	Listens & Records
School 1	A & B C & D	Assessor A Assessor C	Assessor B Assessor D
School 2	A & C B & D	Assessor A Assessor B	Assessor C Assessor D
School 3	A & D B & C	Assessor A Assessor B	Assessor D Assessor C
School 4	B & A D & C	Assessor B Assessor D	Assessor A Assessor C
School 5	C & A D & B	Assessor C Assessor D	Assessor A Assessor B
School 6	D & A C & B	Assessor D Assessor C	Assessor A Assessor B
Etc. …			

Schedule with 5 assessors

School	Assessor Pairing	Assesses & Records	Listens & Records
School 1	A & B	Assessor A	Assessor B
	C & D	Assessor C	Assessor D
	E	Assessor E	-----
School 2	E & A	Assessor E	Assessor A
	B & C	Assessor B	Assessor C
	D	Assessor D	-----
School 3	D & E	Assessor D	Assessor E
	A & C	Assessor A	Assessor C
	B	Assessor B	-----
School 4	C & E	Assessor C	Assessor E
	B & D	Assessor B	Assessor D
	A	Assessor A	-----
School 5	A & D	Assessor A	Assessor D
	E & B	Assessor E	Assessor B
	C	Assessor C	-----
School 6	B & A	Assessor B	Assessor A
	D & C	Assessor D	Assessor C
	E	Assessor E	-----
School 7	A & E	Assessor A	Assessor E
	C & B	Assessor C	Assessor B
	D	Assessor D	-----
School 8	E & D	Assessor E	Assessor D
	C & A	Assessor C	Assessor A
	B	Assessor B	-----
School 9	E & C	Assessor E	Assessor C
	D & B	Assessor D	Assessor B
	A	Assessor A	-----
School 10	D & A	Assessor D	Assessor A
	B & E	Assessor B	Assessor E
	C	Assessor C	----
Etc…			

If the assessment team is made up of an odd number, rotate the teams, leaving one person to skip the inter-assessor reliability for that one day. Create a schedule for assessment, like the ones above. It is critical that this schedule be created to avoid any confusion on the part of the assessors.

For any further questions or clarifications, contact the Department of Education and Child Protection research team at learningassessment@savechildren.org.

ANNEX L: 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?
Treat_phase	Ordinal (1-6)	treat_phase	Set value labels according to project	In which phase of the study is this treatment-school student?

Section: Demographic	Format	Label name	Label values	Variable label
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?

ANNEX M: RECOMMENDATIONS FOR EQUATING

Expert panelists and workshop participants at the 2015 USAID workshop "Improving the Quality of EGRA Data: A Consultation to Inform USAID Guidance on the Administration of Early Grade Reading Assessments" discussed the topic of equating same-language subtasks for multiple forms of an instrument. The equating panel's detailed technical recommendations and areas for further deliberation are provided below.

M.1 Recommendations

1. **For subtasks with few items (e.g., 10–25), pilot multiple, newly developed test forms along with baseline forms.** Then compare item-level statistics across forms and use this information (*p*-values and point biserials) to construct midterm/endline forms that most closely mimic the baseline form statistics. This is a simplified common-persons pre-equating approach using classical test theory (CTT).
2. **Do not apply CTT equating approaches to subtasks with few items.** The rationale is clear for subtasks with 3–5 items (such as listening and reading comprehension), but becomes an area for further deliberation when items remain as few as 10–25. While it may be possible to equate using IRT approaches, these require sample sizes of at least 500–1,000 students for more complex models (i.e., two- or three-parameter models). Rasch models and CTT approaches require similar sample sizes—so the choice between them becomes a matter of meeting assumptions (and the suitability of analyzing item-level data).
3. **For linear data and small samples, use CTT equating methods.** In such cases, common-persons piloting can be used for oral reading fluency (ORF) passages and mean or linear equating approaches can be applied (and chosen based on visual fit, bias, and error). With nonlinear data, the process becomes more complicated. This is more appropriate than item-response theory (IRT) equating procedures, given that the ORF measure yields a total score (without useful item-level data).
4. **Ensure that pilot and operational samples are as similar as possible.** Since many equating approaches for EGRA rely on common-persons or randomly

equivalent samples piloting (particularly for ORF), the pilot sample must be as representative of the operational sample as possible, so that equating adjustments applied to the pilot sample are appropriate for the operational data.

5. When using equipercentile equating for equating nonlinear ORF data, **ensure that the sample has students at all possible score points**—which often requires a larger sample size than is feasible in common-persons piloting for EGRA-based studies.

Exhibit M-1 represents recommendations regarding the EGRA summary variables that can or cannot be equated using traditional equating approaches with small samples. There is still room for discussion on this table, particularly with regard to zero scores, subtasks with between 10 and 25 items, and the percent correct of attempted.

Exhibit M-1. EGRA summary variables to be equated (recommendations)

EGRA subtasks	No. of items	Timed score	Zero score	Score	% score	No. attempted	% correct of attempted
Phonetic / Syl-lable Sounds	~20	IRT - Rasch	No	No	No	No	No
Vocabulary	5–10	IRT - Rasch	No	No	No	No	No
Letter Names	100	Anchor-item	No	Anchor-item	Anchor-item	No	No
Letter Sounds	100	Anchor-item	No	Anchor-item	Anchor-item	No	No
Familiar Words	50	Anchor-item	No	Anchor-item	Anchor-item	No	No
Nonwords	50	Anchor-item	No	Anchor-item	Anchor-item	No	No
Oral Reading fluency	~50	Common-persons (equi-percentile)	No	Common-persons (equiper-centile)	Common-persons (equiper-centile)	No	No
Reading Comprehension	~5	IRT - Rasch*	No	No	No	No	No
Listening Comprehension	~5	IRT - Rasch*	No	No	No	No	No
Dictation	10–15	IRT - Rasch	No	No	No	No	No
Maze	10–15	IRT - Rasch	No	No	No	No	No

* = further investigation is necessary.

Note: These approaches are recommended only if the pilot is thought to follow a distribution similar to that of the full survey (this done by random sampling).

M.2 Areas for Further Deliberation

1. **What approaches can be used for equating reading comprehension subtasks (or other subtasks with as few as 5 items)?** Traditional CTT equating approaches do not work in these circumstances, but stepwise or other nonlinear approaches should be investigated. IRT equating should be investigated further to determine feasibility with the shortest subtasks.

2. **What approaches can be used for equating ORF passages when there is evidence of nonlinear relationships across forms**? There is promising evidence of the appropriateness of both circle-arc and equipercentile equating in these situations, but each still comes with limitations that must be explored further.

3. **What are the trade-offs between CTT and IRT equating approaches?** These issues include technical expertise, sample size, piloting procedures, etc. Ultimately, when item-level data are recorded, Rasch analyses for small samples should be preferred.

4. **How should zero scores be handled during equating?** Should students with zero scores on all assessment forms be excluded from equating calculations (or just those with zero scores on any form)? Is it possible to receive a zero on a particular test form but have an equating adjustment produce a nonzero score for that student? How dependent is the handling of zero scores on the equating method to be applied?

5. **What are the implications of using pilot versus operational data for equating?** In the majority of instances, we are limited to using pilot data for equating ORF, but what are the trade-offs in circumstances where either approach is possible? Since post-equating (i.e., operational data) is likely to provide more reliable equating relationships, is there any reason to rely on pre-equating (i.e., pilot data) when both options are available?

6. **How can equated scores be analyzed?** If grades are equated separately, must they also be analyzed separately (thus negating an overall/combined analysis)? If raw scores are equated, can analyses be conducted on percent correct of attempted items?

7. **Order effects should be investigated for tests in which items are randomly sorted within rows.** Inadvertent grouping of difficult items could have an impact on test scores; this needs to be explored further.

8. **There is a need to explore dimensionality between subtasks.** If a reasonable level of unidimensionality can be demonstrated, then it is possible that equating

from some subtasks can be extrapolated to others. Otherwise, all equating is restricted to the component level (i.e., subtasks), which may limit generalizability with regard to overall reading achievement.

ANNEX N: DETAILED TECHNICAL RECOMMENDATIONS ON PUBLIC-USE FILES

Section 10.6 of this toolkit outlines the steps required before EGRA data are made available to the public. This annex presents additional detailed technical recommendations of the panel on public-use files, as agreed upon at the 2015 USAID workshop "Improving the Quality of EGRA Data: A Consultation to Inform USAID Guidance on the Administration of Early Grade Reading Assessments."

N.1 Specific Recommendations for Cleaning, Finalizing, and De-Identifying Data

N.1.1 Cleaning

1. USAID reinforces the benefits of adopting a master codebook to Evaluation/Implementing Partners.
2. The codebook prepared by the USAID Education Data for Decision Making (EdData II) project[34] is used as the basis for the master codebook. The panel recommends the development of a codebook for the supplementary instruments, such as demographic information in the student questionnaire.
3. Whenever possible, variable names are defined with at most 12 characters and variable labels with at most 80 characters.
4. PUFs are self-describing, with categorical data, using the categories as values instead of a numeric code.
5. In order to avoid unsubstantiated generalizations, data must be removed for all geographical areas that were not used for sampling purposes and have too few schools in the sample to obtain proper precision estimates (such as district, enumeration area, locality, and neighborhood).

[34] The codebook (RTI International, 2014a), in the format of an Excel spreadsheet, is available from the EdData II website: https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=389. Also Annex L of this toolkit presents a sample codebook.

N.1.2 Finalizing

In cases in which a complex sample methodology has been used, whenever possible, the survey specifications should be set in the PUF data file to minimize misspecifications by public users.

In order to mitigate misspecification issues during data analysis by public users, the analysis data set is used as the basis for the PUF. Where possible and appropriate, the researchers merge (e.g., teacher and student data) or append (e.g., baseline and endline) data files to avoid the need for public users to manipulate multiple data files.

N.1.3 De-identifying

All potential personally identifying information must be removed from data sets before they are made publicly available. Below are general recommendations for removing and anonymizing identifiable information.[35]

1. Remove personally identifying information such as name, home address, telephone number, and national identification number
2. Remove school names and names of any other institutions and individuals that may have been collected during the data capture process.
3. Remove all information used to contact and find the schools or institutions (such as address, telephone number, head teacher's name, global positioning system (GPS) coordinates).
4. Data used for sampling purposes might include personally identifying information: Anonymize the data but do not destroy it. It is important to keep a restricted data set for matching the anonymized variable values with the non-anonymized variable values.
5. Anonymize all variables that contain the country's official codes (e.g., school or institutional code, teacher code).

N.2 Dissemination of PUF Data

USAID is expected to make PUF files containing early grade reading assessment data publicly available through the Secondary Analysis for Results Tracking Education portal (SART Ed) and the Development Data Library (DDL). In order to

[35] For a detailed discussion, see Annex A of Optimal Solutions Group (2015).

facilitate data exploration by the public, the panel recommends that in addition to data files, accompanying documentation also be provided.

Along with the PUF, well-documented information that helps the public users to familiarize themselves with the data should also be posted. The following information should be provided to the users:

1. Written data analysis report submitted to and approved by USAID.
2. The questionnaires and assessment tools used to collect the data. (In order not to compromise materials that may be used for future studies within the same project, these items may be provided only after the project is complete.)
3. The background information and all relevant documentation.

In addition to requiring the reports and data collection instruments, USAID reiterates to Evaluation/Implementing Partners the importance of documenting the names and descriptions of the key variables and settings needed for proper data analysis, including:

1. An explicit definition of the population of interest, including the source of the list frame that was used to draw the sample. The documentation indicates the total number of schools and an estimated number of students that the sample is meant to represent. These numbers also match the weighted data estimates. If the survey involves an intervention/control, the numbers are reported out by intervention/control.
2. Variables needed to properly analyze the complex data based on the sample methodology (e.g., for each stage of sampling: all sampled items, the stratification variable, and finite population correction variable, as well as the final weight variable).
3. Variables for the research design (e.g., treatment, year, and cohort if the research design is a staggered impact evaluation assessment.
4. An explanation of sample methodology settings (e.g., variance handling), such that survey design characteristics can be used independent of proprietary software.
5. A complete codebook containing:
    a. A list of all the variables in the data set.
    b. Each variable label and format, as well as value labels (where relevant).

c. A formal description of the computation used to generate calculated variables (e.g., oral reading fluency).

d. Total number of observations in the data set.

While USAID's planned data repositories are under construction, Implementing Partners and Evaluators make their PUFs with early grade reading assessment data publicly available.

1. Post the PUF in an accessible location online, accompanied by its documentation as specified above (i.e., all items are located in one zipped file or the website contains links to these documents).

2. Create the PUF using a nonproprietary data file, and when possible a proprietary data file.

3. For the nonproprietary file, create a csv, comma-separated values text file.

4. For the nonproprietary file, create either a Stata .dta file and/or an SPSS .sav file (along with the SPSS.csaplan file).

The panel also encourages USAID to consider the development of guidelines for evaluation reports based on early grade reading reports, similar to USAID's general guidance on preparing evaluation reports (USAID, 2012), as this would make it easier for the general public to locate information in the reports. It would also guarantee that the same basic information could be found across reports.

References for Annex N

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. Retrieved from https://sartdatacollection.org/images/SARTDataSharingManualFeb2015.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

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. Retrieved from https://www.usaid.gov/sites/default/files/documents/1870/How-to-Note_Preparing-Evaluation-Reports.pdf

ANNEX O: EGRA DATA ANALYSIS

For every pupil score estimate reported, a visual of the score distribution, such as the ones shown in **Exhibits O-1 through O-3**, must be graphically presented to support the reader's interpretation of the estimate provided.

Exhibit O-1. Example of difference-in-difference analysis

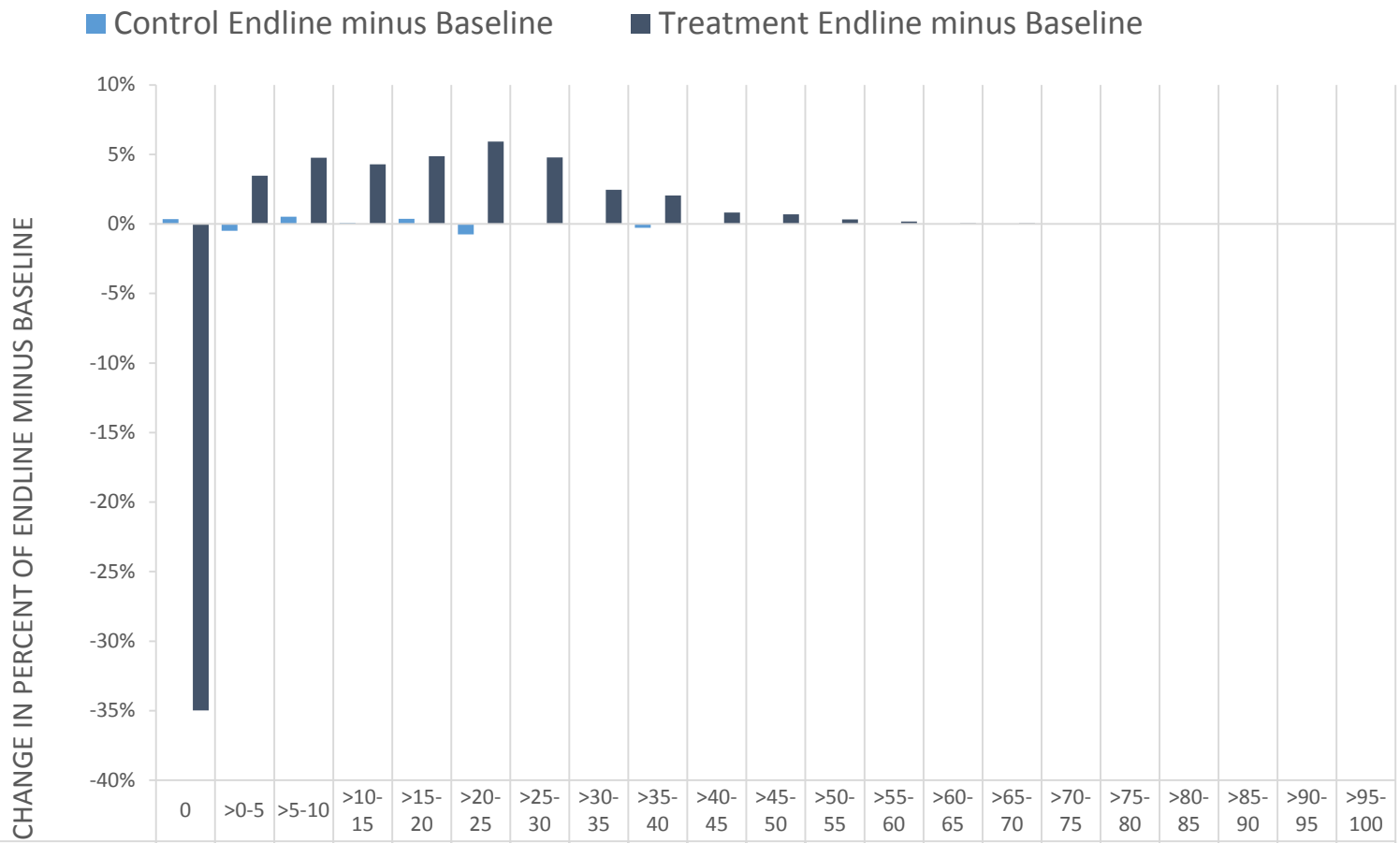
	Baseline					Endline							
Treatment	**Mean fluency (wpm)**	**Standard error**	**Number of sampled students**	***t*-stat**	***p*-value**	**Mean fluency (wpm)**	**Standard error**	**Number of sampled students**	***t*-stat**	***p*-value**	**Differ-ence–in–differ-ence**	***p*-value (DID)**	**Effect size**
Control	4.5	0.6	656	–	–	9.5	1.6	475	–	–	–	–	–
Intervention	5.2	1.2	349	0.510	0.611	11.7	1.1	480	1.189	0.236	1.5	0.490	0.12

Calculations:

Difference-in-Difference: (Mean endline treatment – mean baseline treatment) – (mean endline control – mean baseline control)

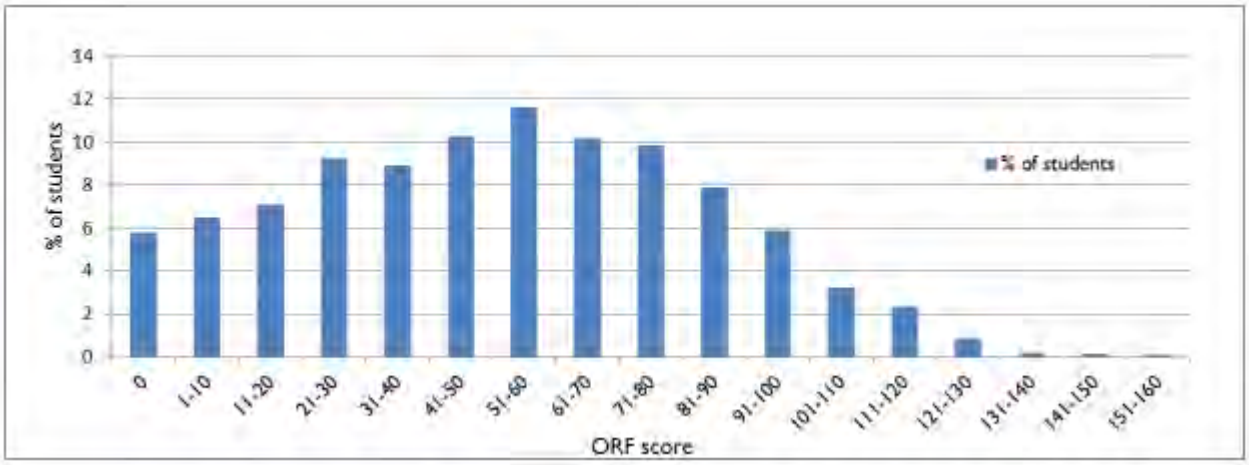
Effect Size (Cohen's *d*): Difference-in-difference / pooled standard deviation

Exhibit O-2. Example of distributional comparison of differences between control and treatment



	0	>0-5	>5-10	>10-15	>15-20	>20-25	>25-30	>30-35	>35-40	>40-45	>45-50	>50-55	>55-60	>60-65	>65-70	>70-75	>75-80	>80-85	>85-90	>90-95	>95-100
Control Endline minus Baseline	0%	0%	1%	0%	0%	-1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Treatment Endline minus Baseline	-35%	3%	5%	4%	5%	6%	5%	2%	2%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Exhibit O-3. Oral reading fluency (ORF) distribution – Indonesia, 2013



Source: Stern, J. & Nordstrum, L. (2014). *Indonesia 2014: The National Early Grade Reading Assessment (EGRA) and Snapshot of School Management Effectiveness (SSME) surv*ey. Prepared for USAID/Indonesia under the Education Data for Decision Making (EdData II) project, Task Order No. AID-497-BC-13-00009 (RTI Task 23). Research Triangle Park, NC: RTI International. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=680

ANNEX P: ENGLISH ORAL READING FLUENCY NORMS

Hasbrouck & Tindal Oral Reading Fluency Data

This table shows the oral reading fluency rates of students in grades 1 through 8, based on an extensive study conducted by Jan Hasbrouck and Gerald Tindal. The results of their study are published in a technical report entitled, "Oral Reading Fluency: 90 Years of Measurement," which is available on these websites:

- **ERIC website:** eric.ed.gov/?id=ED531458
- **BRT website:** www.brtprojects.org/publications/technical-reports

This table can help you assess the oral reading fluency of your students relative to their peers. Students scoring 10 or more words below the 50th percentile using the average score of two unpracticed readings from grade-level materials need a fluency-building program. Teachers can also use the table to set long-term fluency goals for struggling readers.

For more information:

- **Essential Components of Reading:** readnaturally.com/components
- **Correlation Between Oral Reading Fluency and Overall Reading Achievement:** readnaturally.com/correlation
- **Read Naturally Tools for Assessing Fluency:** readnaturally.com/assessment-tools
- **Read Naturally Intervention Programs That Develop Fluency:** readnaturally.com/fluency-interventions

Grade	Percentile	Fall WCPM*	Winter WCPM*	Spring WCPM*	Avg. Weekly Improvement**
1	90		81	111	1.9
	75		47	82	2.2
	50		23	53	1.9
	25		12	28	1.0
	10		6	15	0.6
2	90	106	125	142	1.1
	75	79	100	117	1.2
	50	51	72	89	1.2
	25	25	42	61	1.1
	10	11	18	31	0.6
3	90	128	146	162	1.1
	75	99	120	137	1.2
	50	71	92	107	1.1
	25	44	62	78	1.1
	10	21	36	48	0.8
4	90	145	166	180	1.1
	75	119	139	152	1.0
	50	94	112	123	0.9
	25	68	87	98	0.9
	10	45	61	72	0.8
5	90	166	182	194	0.9
	75	139	156	168	0.9
	50	110	127	139	0.9
	25	85	99	109	0.8
	10	61	74	83	0.7
6	90	177	195	204	0.8
	75	153	167	177	0.8
	50	127	140	150	0.7
	25	98	111	122	0.8
	10	68	82	93	0.8
7	90	180	192	202	0.7
	75	156	165	177	0.7
	50	128	136	150	0.7
	25	102	109	123	0.7
	10	79	88	98	0.6
8	90	185	199	199	0.4
	75	161	173	177	0.5
	50	133	146	151	0.6
	25	106	115	124	0.6
	10	77	84	97	0.6

*WCPM = Words Correct Per Minute

www.readnaturally.com

**Average words per week growth

United States Agency for International Development
Office of Education
Bureau for Economic Growth, Education, and Environment (E3)
1300 Pennsylvania Avenue, N.W.
Washington, DC 20523, USA
www.USAID.gov