EARLY GRADE READING ASSESSMENT TOOLKIT

March 30, 2009

Prepared for
The World Bank
Office of Human Development

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

Table of Contents

Exhibits

Abbreviations

AKF	Aga Khan Foundation
ANOVA	analysis of variance
CIASES	Centro de Investigación y Acción Educativa Social [Nicaraguan NGO]
CLPM	correct letters per minute
CTOPP	Comprehensive Test of Phonological Processing
CVC	consonant-vowel-consonant
CWPM	correct words per minute
DEFF	design effect
DFID	British Department for International Development
DIBELS	Dynamic Indicators of Basic Early Literacy
EFA	[United Nations] Education for All [campaign]
EGR	Early Grade Reading [project, Kenya]
EGRA	Early Grade Reading Assessment
EMACK	Education for Marginalized Children in Kenya
GPC	grapheme-phoneme correspondence
ICC	intraclass correlation coefficient
LAMP	Literacy Assessment and Monitoring Programme [UNESCO]
LCD	liquid-crystal display
LQAS	Lot Quality Assurance Sampling
MDG	Millennium Development Goal
NGO	nongovernmental organization
PASEC	Programme d'Analyse des Systems Educatifs de la Confemen
PGC	phoneme-grapheme correspondence
PISA	[Organisation for Economic Co-Operation and Development's] Programme for International Student Assessment
PPVT	Peabody Picture Vocabulary Test
SAQMEC	Southern Africa Consortium for the Measurement of Educational Quality
SD	standard deviation
SE	standard error
TIMSS	Trends in International Mathematics and Science Study
TOPA	Test of Phonological Awareness
UNESCO	United Nations Educational, Scientific and Cultural Organisation
USAID	United States Agency for International Development

Glossary of Terms

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

Automaticity/Fluency. The bridge between decoding and comprehension. Fluency in word recognition so that the reader is no longer aware of or needs to concentrate on the mental effort of translating letters to sounds and forming sounds into words. At that point, the reader is decoding quickly enough to be able to focus on comprehension.

Blend. A group of two or more consecutive consonants that begin a syllable (as *gr-* or *pl-* in English).

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

Digraph. A group of consecutive letters whose phonetic value is a single sound (e.g., *ea* in *bread, ch* in *chin*). Some digraphs are graphemes (see below).

Floor effect. A statistical term to denote an artificial lower limit on the possible values for a variable, causing the distribution of scores to be skewed. For example, the distribution of scores on an EGRA ability test will be skewed by a floor effect if the test is much too difficult for most children in the early grades to perform at a sufficient skill level to allow for analysis.

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

Grapheme. The most basic unit in an alphabetic written system. Graphemes combine to create phonemes (see Phoneme). 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.

Logograph. A single grapheme that also forms a word or morpheme; for example, “a” in Spanish or “I” in English.

Morpheme. Smallest linguistic unit with meaning. Different from a word, as words can be made up of several morphemes (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-).

Morphograph. Smallest unit of meaning in a word.

Metaphonology. See Phonological awareness.

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).

Orthographic. The art of writing words with the proper letters according to usage; spelling.

Phoneme. The smallest linguistically distinctive unit of sound allowing for differentiation of two words within a specific language (e.g., “top” and “mop” differ by only one phoneme, but the meaning changes).

Phonological awareness. A general appreciation of the sound structure of language, as demonstrated by the awareness of sounds at three levels of structure: syllables, onsets and rimes, and phonemes.

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) and “oil” (the rime).

Acknowledgments

This 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 the contributions to the development and proliferation of the Early Grade Reading Assessment (EGRA), special thanks are extended to Helen Abadzi, Marilyn Jager Adams, Rebecca Adams, Rukmini Banerji, Danielle Bechenec, Penelope Bender, Sandra Bertoli, Joanne Capper, Vanessa Castro, Colette Chabbott, Madhav Chavan, Marguerite Clarke, Penny Collins, Luis Crouch, Marcia Davidson, Joe DeStefano, Maria Diarra, Zakeya El-Nahas, Deborah Fredo, Ward Heneveld, Robin Horn, Sandra Hollingsworth, Matthew Jukes, Cheryl Kim, Medina Korda, José Ramon Laguna, Nathalie Lahire, Sylvia Linan-Thompson, Corrine McComb, Emily Miksic, Amy Mulcahy-Dunn, Lily Mulatu, Lynn Murphy, Robert Prouty, Alastair Rodd, Mike Royer, Momar Sambe, Ernesto Schiefelbein, Dana Schmidt, Philip Seymour, Linda Siegel, Jennifer Spratt, Liliane Sprenger-Charolles, Helen Stannard, Jim Stevens, Ian Smythe, Sana Tibi, Gerald Tindal, Palesa Tyobeka, Dan Wagner, and Jim Wile.

Extensive peer review comments for this toolkit and suggestions for instrument development were provided by Marcia Davidson, Sandra Hollingsworth, Sylvia Linan-Thompson, and Liliane Sprenger-Charolles.

Development of EGRA would not have been possible without the support of the nongovernmental organization 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 is responsible for primary authorship of the toolkit, with contributions from Luis Crouch, Amy Mulcahy-Dunn, and Marguerite Clarke; editing support was provided by Erin Newton. 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 or the World Bank. Please direct questions or comments to Amber Gove at agove@rti.org.

I. Introduction

Why Focus on Early Grade Reading?

Countries around the world have boosted primary school enrollment to historically unprecedented rates. Seeking to honor the commitments of the United Nations' Education for All (EFA) campaign and Millennium Development Goals (MDGs), low-income countries, with international support, are enrolling children in primary school at nearly the rates of high-income countries. But are students learning?

The evidence, when available, indicates that average student learning in most low-income countries is quite low. A recent evaluation of World Bank education lending shows that improvements in student learning are lagging significantly behind improvements in access to schooling (World Bank: Independent Evaluation Group 2006). Results from those few low-income countries that participate in international assessments such as PISA or TIMSS (and inferring from the results of regional assessments such as PASEC and SACMEQ)[1] indicate that the median child in a low-income country performs at about the 3rd percentile of a high-income country distribution (i.e., worse than 97 percent of students who were tested in the high-income country).[2] From these results, we can tell what low-income country students did *not* know, but cannot ascertain what they *did* know (often because they scored so poorly that the test could not pinpoint their location on the knowledge continuum). Furthermore, most national and international assessments are paper-and-pencil tests administered to students in grade 4 and above (that is, they assume students can read and write). It is not always possible to tell from the results of these tests whether students score poorly because they lack the knowledge tested by the assessments, or because they lack basic reading and comprehension skills.

The ability to read and understand a simple text is one of the most fundamental skills a child can learn. Without basic literacy there is little chance that a child can escape the intergenerational cycle of poverty. Yet in many countries, students enrolled in school for as many as 6 years are unable to read and understand a simple text. Recent evidence indicates that learning to read both *early* and at a sufficient *rate* (with comprehension) is essential for learning to read well. 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 and eventually drop out. Global efforts to expand access to education may be undermined if parents, faced with difficult economic choices and the knowledge that students are not acquiring basic reading skills, remove their children from school. In many countries it is apparent that this trend may already be occurring: while more students are enrolled, primary school completion and cohort survival rates (a measure of education system output as well as student "survival" in the system) have not kept pace with expanded enrollments.

[1] Organisation for Economic Co-Operation and Development's Programme for International Student Assessment (PISA); Trends in International Mathematics and Science Study (TIMSS); Programme d'Analyse des Systems Educatifs de la Confemen (PASEC); Southern Africa Consortium for the Measurement of Educational Quality (SACMEQ).

[2] See, for example, the percentile breakdowns in table D.1. in Mullins, Martin, Gonzalez & Chrostowski (2004). Similar conclusions can be derived from the OECD PISA Report (2004), table 2.1.c, for example. Typically only middle-income countries participate in these international assessments. By looking at the few poor countries that do participate in these assessments, and by linking these and the middle-income ones that do participate, to regional assessments such as PASEC and SACMEQ, we can extrapolate that in poor countries the median child must be achieving at about the 3rd percentile of the developed country distribution (Crouch & Winkler, 2007). For example, mean grade 8 performance in Ghana on TIMSS 2003 was 274, but average 5th-percentile performance across high-income countries was 376. In a few of the more advanced middle-income countries, such as Brazil or Tunisia, the mean performance can be above the 5th percentile of the high-income countries.

Measuring Learning: The Instrument and the Toolkit

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

> "In some countries, 50 percent of fourth-grade students do not understand the meaning of the texts they read (in one public school class, I found 20 non-reading students in a class of 29), but the majority of these students attend schools that cater to families in the 'lower half of the income bracket.' This means that 90 percent of the students in this half of the population do not understand what they read (even though many complete their primary schooling). In this situation a good literacy program (in the first two grades of primary school) can have an immense impact on the performance of the education system."
>
> —Ernesto Schiefelbein, Former Minister of Education, Chile

To respond to this demand, work began on the creation of an Early Grade Reading Assessment (EGRA). What was needed was 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: recognizing letters of the alphabet, reading simple words, and understanding sentences and paragraphs. Development of EGRA began in October 2006, when USAID, through its EdData II project, contracted RTI International to develop an instrument for assessing early grade reading. The objective was to help USAID partner countries begin the process of measuring, in a systematic way, how well children in the early grades of primary school are acquiring reading skills, and ultimately to spur more effective efforts to improve performance in this core learning skill.

Based on a review of research and existing reading tools and assessments, RTI developed a protocol for an individual oral assessment of students' foundation reading skills. To obtain feedback on this protocol and to confirm the validity of the overall approach, RTI convened a meeting of cognitive scientists, early-grade reading instruction experts, research methodologists, and assessment experts to review the proposed key components of the instrument. During the workshop, participants were charged with bridging the gap between research and practice—that is, merging advances in the reading literature and cognitive science research fields with assessment practices around the world. Researchers and practitioners presented evidence on how to measure reading acquisition within the early primary grades. In addition, they were asked to identify the key issues to consider in designing a multicountry, multilanguage, early grade reading assessment protocol. The workshop, hosted by USAID, the World Bank, and RTI in November 2006, included 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. A summary of the workshop proceedings can be found at www.eddataglobal.org under News and Events.

During 2007, the World Bank supported an application of the draft instrument in Senegal (French and Wolof) and The Gambia (English), while USAID supported the application in Nicaragua (Spanish). In addition, national governments, USAID missions, and nongovernmental organizations (NGOs) in South Africa, Kenya, Haiti, Afghanistan, Bangladesh, and other

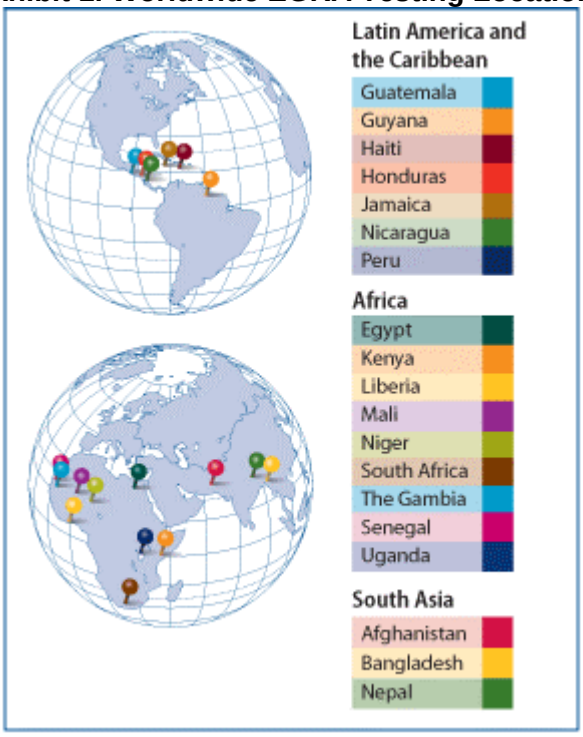
countries began to experiment with the application of certain components of the assessment (with and without the involvement of RTI). In the interest of consolidating these experiences and developing a reasonably standardized approach to assessing children's early reading acquisition, the World Bank requested that RTI develop a "toolkit" which would guide countries beginning to work with EGRA in such areas as local adaptation of the instrument, fieldwork, and analysis of results.

The objective of this document is to provide practical guidance to ministries of education and their partner agencies to support the application of EGRA in English. Nonetheless, occasional references to development of EGRA in other languages are provided for illustrative purposes.

Worldwide Applications

The EGRA instrument presented here and the results obtained through field testing have generated considerable discussion and interest within the donor community and among country ministries of education. Based on the results of the EGRA application in their countries, Ministry staff from The Gambia and South Africa developed detailed teacher handbooks for instructional approaches and trained teachers in their use. International NGOs also began to use the draft instruments in their work in the developing world. Plan International, working in francophone Africa, developed teacher training and instructional approaches and has piloted them in a number of schools in Mali and Niger. Save the Children used the French version developed for Senegal and adapted it for use in Haiti in Haitian Creole, and has plans for application in a number of other countries.

Exhibit 1. Worldwide EGRA Testing Locations



Other experiments with EGRA have added additional rigor to the model. In Kenya, RTI and the Aga Khan Foundation are applying baseline and follow-on assessments in treatment and control schools. In each of the 20 treatment schools, teachers are being trained in early reading instruction techniques and continuous assessment. Recently a version adapted for Arabic in Egypt was successfully applied in 60 schools. Complementary efforts are under way in several other countries (see Exhibit 1 map, up to date as of February 2009). (For a list of countries working with EGRA, please see the Documents and Data link on the www.eddataglobal.org website).

One of the reasons for the high level of interest in the EGRA tool is the direct link to advances in both reading and cognitive development research. While much of this research stems from working with children in high-income countries, the basis for such research lies with advances in neuroscience and thus has relevant lessons for low-income countries.[3] The importance of "early" should also be

[3] For a reasonably accessible summary of advances in neuroscience and cognitive development research, see Abadzi (2006).

emphasized here: evidence from several countries indicates the presence of what Stanovich (1986) tags as a "Matthew effect" in reading acquisition.[4] That is, if strong foundation skills are not acquired early on, gaps in learning outcomes (between the "haves" and the "have-nots") grow larger over time.

A second reason is the intuitive simplicity of the measure for Ministry staff, teachers, and parents. Most people would agree that regardless of instructional approach, children enrolled in school for 3 years should be able to read and comprehend a simple text. The experience of the Indian NGO Pratham and joint efforts of the British Department for International Development (DFID) and the World Bank in Peru have shown that simple but reasonably rigorous measures of early reading can have a substantial impact on the national dialogue surrounding school quality and student learning.[5]

Finally, EGRA is designed to be a method-independent approach to assessment: It doesn't matter how reading is being taught—research shows that the skills tested in EGRA are necessary but not sufficient for students to become successful readers. The reading "wars" are alive and well in many low-income countries, often miring ministries of education and teaching centers in seemingly endless debates between the "whole-language" and "phonics-based" approaches. Nonetheless, the evidence for reading acquisition in English points to a comprehensive approach, based on five essential components identified by the U.S. National Reading Panel (National Institute of Child Health and Human Development, 2000): phonics, phonemic awareness, vocabulary, fluency, and comprehension. EGRA utilizes each of these components, emphasizing the foundation skills of reading acquisition.

Ideas on how to continue to improve and use (or not use) EGRA are evolving. In March 2008, nearly 200 participants from some 40 countries attended a 3-day workshop in Washington, DC. Participants included representatives from donor organizations and foundations, Ministry and NGO staff, and international reading assessment and instruction experts. The primary objectives of the workshop were twofold. First was to continue to generate interest in, and awareness of, EGRA activities among the donor community and potential participant countries. Second was to prepare a select group of interested countries for starting actual applications of EGRA. Participants gained technical knowledge of the importance of early reading instruction and assessment; an understanding of the research underpinning EGRA, including the foundation steps to reading acquisition; and potential uses for information generated by EGRA. For more information on the workshop and video links of the presentations, please see http://go.worldbank.org/0SFS7PP330.

Toolkit Audience and Contents

This toolkit is divided into seven sections and is intended for use by Ministry of Education staff and professionals in the field of education development. More specific audiences for certain sections of the toolkit are noted as appropriate.

[4] The term "Matthew effect," often used in the context of reading research and summarized as "the rich get richer and the poor get poorer," derives from a statement that appears in a biblical parable in the book of Matthew: "For to all those who have, more will be given, and they will have an abundance; but from those who have nothing, even what they have will be taken away" (25:29).

[5] Pratham's Annual Status of Education Report (2005) documents results from a simple reading and math assessment administered to 330,000 children in 10,000 villages using an all-volunteer staff. For the report and more information, see www.pratham.org. In Peru, World Bank and DFID efforts led to the inclusion of school quality and early reading issues in the national presidential debate. A link to a video developed for the purposes of policy dialogue can be found at www.eddataglobal.org (main page). For additional information on the Peru assessment and results see Abadzi, Crouch, Echegaray, Paco & Sampe (2005).

The document seeks to summarize a large body of research in an accessible manner, while providing practical, detailed tips for designing and conducting a sample-based, baseline EGRA to raise awareness and promote policy dialogue.

The toolkit is not intended to be a comprehensive review of all reading research. In the interest of brevity and understanding, the toolkit does not cover all aspects of, and alternatives to, reading assessment. It should also be noted that EGRA is a work in progress; readers of this toolkit should check the EdData website for the latest instruments and updates. Furthermore, EGRA is not an "off-the-shelf" assessment—each new country application requires review of vocabulary and development of context-appropriate reading passages. Development of EGRA in local languages, especially when centrally located word/vocabulary lists are not widely available, requires considerable effort and should be done in conjunction with an expert in the local language (direct translation is not appropriate, as discussed in detail below). Finally, EGRA is designed to complement, rather than replace, existing curriculum-based pencil-and-paper assessments.

Following this introduction, Section II is an overview of the purposes and uses of the assessment. Section III includes the conceptual framework and research foundations (the theoretical underpinnings of the assessment). Section IV discusses preparatory steps to administration of the assessment, including a design workshop for construction of the EGRA instrument. Section V advises on sample selection, training of EGRA enumerators, realities that country teams can expect in the field, and means for collecting the data. Section VI is an overview of analyses to be conducted. Section VII provides guidance on interpretation of results and some summary implications for policy dialogue related to improving instruction and reporting results to schools.

Annexes include sample oral reading fluency norms for English by grade; a sample EGRA workshop schedule developed for Kenya; and sample teacher lesson plans based on EGRA results. In addition, an example of how EGRA can influence education Ministry policy is represented by a letter from the Director for Education in South Africa. Finally, technical annexes describing sample size considerations and tests of reliability are included.

As noted earlier, this toolkit was created to inform the development and use of EGRA in English, with occasional brief notations about usage in other languages. Separate versions for French and Spanish (with appropriate literature specific to those languages) have been developed with USAID financing.

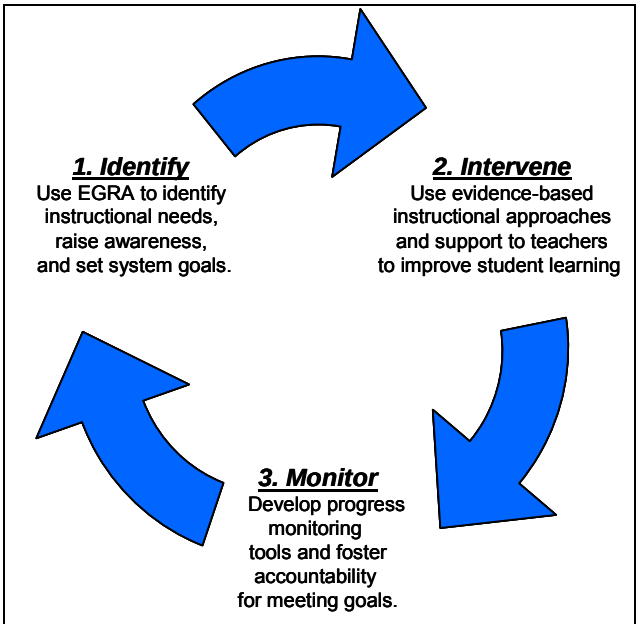
II. Purpose and Uses of EGRA

Although it was clear from the outset that EGRA would focus on the early grades and the foundation skills of reading, the uses to which the results should be put were more open to debate. Interested donors pushed for cross-country comparability and system-level measures that could report on the effectiveness of their investments. Ministries requested an instrument that could tell them how to support teachers through additional training and other means. Finally, teachers required a tool that could help them identify individual children who needed additional help while also assessing the effectiveness of their own instruction. Could a single instrument be designed to meet all of these needs?

The answer is "no." The EGRA instrument as developed and explained in this toolkit is designed to be a sample-based "system diagnostic" measure. Its purpose is to document student performance on early grade reading skills in order to inform ministries and donors regarding system needs for improving instruction. To be clear, as it is currently designed, EGRA is not intended for direct use by teachers, nor is it meant to screen individual students. It is also most certainly not intended to be a high-stakes accountability measure for making funding decisions or determining student grade passing. But that does not mean that the development of one version of the instrument cannot inform another version (with a different purpose and use). The measures included in this version of EGRA could be adapted for teacher use in screening of individual students and, with multiple forms that are equated, EGRA could be used to monitor progress of students within a given instructional program. These alternate uses are only tangentially discussed in the current toolkit.

The EGRA subtest measures—including letter recognition, nonsense word decoding, and oral reading fluency—have been used to fulfill a diverse range of assessment needs, including screening, diagnostic, and progress-monitoring purposes. Using results on oral reading fluency from thousands of students across the United States (see Annex A), education practitioners and researchers have screened students for learning difficulties, diagnosed student strengths and weaknesses to guide instruction, and made decisions regarding the effectiveness of their teacher training and professional development programs. In each of these cases, the instrument (and sampling scheme) **must be** adapted to reflect the purpose of the assessment (a critical aspect to consider in constructing and using any assessment tool) (Hasbrouck & Tindal, 2006; Kame'enui et al., 2006; Kaminski et al., 2006). Design implications for each of these additional approaches and required modifications to EGRA are discussed briefly, below.

Exhibit 2. The Continuous Cycle of Improving Student Learning



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.

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 2 above, 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. General benchmarking and creation of goals for future applications can also be done during the initial EGRA application. Based on the results, education ministries or local systems can then *intervene* to modify existing programs using evidence-based instructional approaches to support teachers for improving foundation skills in reading. Results from EGRA can thus inform the design of both pre-service and in-service teacher training programs.

Once these 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 teacher and education administrator responsibility for making sure students make progress in achieving foundation skills.

Using EGRA to Identify System Needs

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

First, in many high-income countries, teachers (and system administrators) can look to existing national distributions and performance standards for understanding how their students are performing compared to others. By comparing subgroup student performance in relation to national distributions and performance standards, system administrators in the United States and Europe 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.
The problem for low-income countries is that similar benchmarks based on locally generated results are not (yet) available. Building on the goals for correct words per minute (CWPM) developed by researchers and teachers for use in a number of countries, including Chile, Spain, and the United States (and as noted in Crouch, 2006; World Bank: Independent Evaluation Group, 2006), it may be possible to derive some estimates for English and Spanish, in order to make broad comparisons. The implications of this are discussed in detail in Section VII. At this time we suggest that countries work to build their benchmarks during the process of applying EGRA.

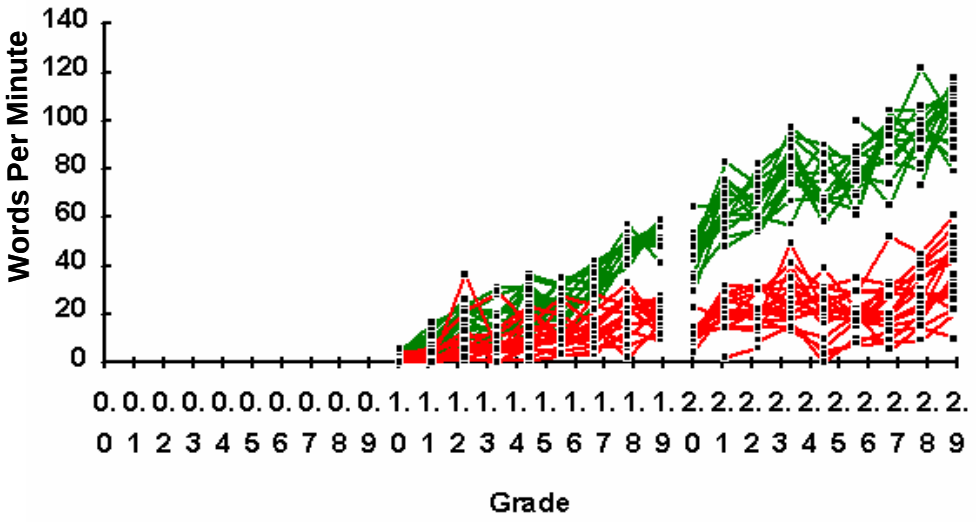
Second, for a measure to be useful for diagnosing early reading difficulties, it must be correlated with later desired reading skills. For example, the ability to recite the names of all the presidents of the United States might be a useful skill, but is likely not correlated with reading skills. Just as a doctor would not measure the length of a patient's foot to determine whether the patient is predisposed to cancer later in life, we would not want to diagnose problems in reading

performance based on a measure that was not related to subsequent performance or student learning outcomes.

Even without research for low-income countries, the wide use of EGRA subtests, or tasks, in other countries as predictive measures can guide the use of EGRA in low-income country contexts. That is, despite the lack of low-income country results at this time, we know enough about the ability of these measures to predict later reading performance and higher-level skills that we can say with reasonable confidence that the predictive aspect of the EGRA tasks (including letter identification, word reading, and others—see Section IV) should function in much the same way in low-income countries. As an example of the predictive power of the tasks that comprise EGRA and similar tools, oral reading fluency has been shown to be predictive of later skills in reading and comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001). The importance of fluency as a predictive measure does, however, decline over time. As students become more proficient and reading comes automatically to them, vocabulary becomes a much more important predictor of later academic success (Yovanoff, Duesbery, Alonzo, & Tindall, 2005).

Third, it makes little sense to measure something that we have no hope of changing through additional instruction. EGRA is valuable as a diagnostic tool precisely because it includes measures of those skills that can be improved (accompanied by teacher support for instruction). Exhibit 3 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 tailored instruction for reading improvement. The green lines in the upper part of the graph show monthly results for students who could read at least 40 words per minute at the end of first grade, while the red lines are the results for students who read less than 40 words per minute at the end of first grade (each unit on the horizontal axis represents a month in the school year).

Exhibit 3. Student Words-per-Minute Scores, Grades 1 and 2



Source: Good, R. H., III, Simmons, D. C., & Smith, S. B. (1998). Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills. *School Psychology Review, 27*(1), 45-56.
Note: Numbers on the horizontal axis refer to the grade (top row) and month (bottom row).

As can be noted in Exhibit 3, in the absence of additional intervention and remediation, the gap between early skilled readers and less proficient readers increases dramatically toward the end of first grade (and continues to widen over time). Thus, the use of EGRA for motivating the need for effective instructional approaches is clear. If teachers and administrators do not intervene, initial gaps in reading acquisition are likely to increase over time. Although it is beyond the scope of this toolkit to review strategies for improving instruction in each of the areas identified by EGRA, teacher guides and lesson plans are available from a number of sources (see Section VII, Changing Reading Instruction).

Additional Uses of EGRA (with Modifications)

Screening

At the classroom level, a modified version of EGRA could be used to indicate which students are having difficulty, and to enable teachers to change the trajectory of student outcomes through specific intervention in needed areas. For EGRA to be used at the classroom level for individual screening of students, three important changes would need to be made to EGRA.

First, the sampling scheme described in detail in Annex B would be eliminated, as each teacher would apply the EGRA instrument to each student (and identify students with difficulties and devise approaches for remediation).

Second, EGRA would need to be divided into subtests or subtasks, to be administered (or not) depending on the teacher's assessment of each individual student's skills and grade level. EGRA currently includes eight subtasks and is designed to be applied to a sample of students in grades 1–3 at the end of the school year (or grades 2–4 at the beginning of the school year). For ease of application and as EGRA is conducted by enumerators who do not know the individual students (and therefore do not have prior knowledge of student performance), the current design of EGRA requires that all students attempt all portions of the assessment. In a teacher-administered instrument, only some of these tasks would need to be applied for each student. For example, teachers might begin testing students with the words task, and if the student is successful, move on to the connected-text portion of the assessment. Alternatively, if a student cannot read letters, then oral comprehension could be tested and the remainder of the test would not need to be conducted. That said, developers should be judicious in deciding which measures are used; in several pilot countries, children as late as third grade had not mastered the most basic foundation skills (such as letter identification).[6]

Third, a teacher-administered EGRA would require teacher training in the administration and interpretation of EGRA tasks, continuous assessment methods, development of instructional approaches for remediation, and strategies for working with students with reading difficulties. Use of a centrally designed EGRA instrument could be obviated through training of teachers to develop their own simple assessments of student reading skills.

Experiments in Kenya, Mali, and Niger, using explicit training of teachers in how to use a modified version of EGRA to understand how students are performing, accompanied by detailed, tightly sequenced lesson plans, have resulted in significant improvements in both

[6] While EGRA as currently designed is meant to be used with children in the early grades, assessments of oral reading fluency have been used in the United States and elsewhere as late as grade 6. Oral fluency in later grades should be conducted with longer texts than those included in EGRA (for additional information on oral reading fluency in later grades please see Espin & Foegen, 1996; Espin & Tindal, 1998).

teacher practice and student reading outcomes. For an example of a teacher-based approach applied in English, please see the materials developed for Kenya at www.eddataglobal.org (Documents and Data>Kenya). Mali and Niger materials are available from Plan International.

Evaluation of Interventions

For EGRA to be useful for evaluating a given intervention, it must be designed such that the students can be tested multiple times using parallel instruments (to avoid memorization or learning of the instrument passages). Currently, such an approach is being tested in Liberia (with parallel forms developed and piloted in the same schools with plans for use in multiple years). Students tested at the outset of an intervention and subsequent to the instructional intervention would be expected to demonstrate improvement greater than that of a similar group of control students. The interventions to be evaluated should be related to reading and learning outcomes; it would make little sense to evaluate a program that cannot be theorized to be related to improving reading outcomes. For example, a program that seeks to improve school-based management would likely not be able to demonstrate improvements in reading due to the absence of a direct causal link. In other words, EGRA can be used for assessing such a program, but such evaluations likely will not reveal demonstrated improvements in reading outcomes. That is, EGRA should only be used to evaluate programs that seek to improve reading instruction and learning outcomes in the primary grades.

How EGRA Should NOT Be Used

EGRA Is Not a High-Stakes Accountability Tool

EGRA should not be used for high-stakes accountability, whether of a punitive, interventionist, or prize-giving variety. Instead, EGRA should be seen as a diagnostic tool whose main clients and users are Ministry staff, with some possible use in more diffuse forms of social mobilization. The reality is that once an instrument of this sort starts to be used, and if communities are empowered with the knowledge that a common-sense understanding of reading matches fairly well with a science-based instrument, it is inevitable, and desirable, that parents and communities should become involved in monitoring reading progress, in a sort of “softer” or community-based accountability. It is also inevitable that officials, all the way up to the Minister of Education, should develop some interest in knowing how well children are performing, and that some need to report would arise. To reiterate, however: Use of EGRA for formulaic and high-stakes forms of accountability such as prize schemes for teachers should be avoided. EGRA, as it is currently designed (a system-level diagnostic), should not identify students or teachers for subsequent follow-up—with the only possible exception being confidential tracking of students and teachers for purposes of screening and project evaluations.

EGRA Is Not Suited for Cross-Language Comparisons

The issue of comparability across countries is challenging from an assessment perspective. Although all of the country-specific assessments developed to date appear quite similar, even across languages, and would at face value appear to be comparable, differences in language structure and rate of acquisition discourage direct comparisons. 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, Aro, & Erskine, 2003). Regardless of language, all children who learn to read advance from being nonreaders (unable to read words) to partial readers (can read some items but not others) to readers (can read all or a majority of items). In languages with transparent or “shallow” orthographies (often called phonetically spelled languages), the progression through these levels is very rapid (just a few

months of learning); in languages with more complex or "deeper" orthographies, this process can take several years. In English, for example, completing the foundation steps requires two or more years, with a rate of gain of only a few new items per month of learning; in comparison, regular and transparent languages such as Italian, Finnish, and Greek require only about a year of instruction for students to reach a comparable level (Seymour et al., 2003).

Thus, EGRA should not be used to compare results across languages. As languages have different levels of orthographic transparency, it would be unfair to say that Country A (in which all children are reading with automaticity by grade 2) is outperforming Country B (where children reach this level only by grade 3), if Country A's language has a far more transparent orthography than Country B's language. Nonetheless, finding out at which grade children are typically "breaking through" to reading in various countries, and comparing these grades, will be a useful analytical and policy exercise, as long as it is not used for "rankings" or "league tables" or for the establishment of a single universal standard for, say, reading fluency or automaticity. Thus, if a country's population speaks a language with transparent orthography, and automaticity is being acquired two grades later than in countries with similarly orthographically transparent languages, this should be a matter for analysis and discussion. Fine-tuning expectations and goals for different countries, but within the same language, is part of the purpose of EGRA, and is something that will likely gain attention worldwide. Indeed, some early attempts are already being carried out not just to measure but also to begin to establish some simple standards (World Bank, 2007). At this stage, we are still exploring the implications of cross-country (but within-language) comparisons, which will require additional research and debate.

Within the same language group, the challenge in making cross-country comparisons is not so much a question of comparable instrument development (an important, but surmountable difficulty), but differences in local dialects and vocabulary. For example, instruments for Peru and Nicaragua were developed using virtually the same version of the instrument (with minor local modifications in administrator instructions). Comparison of these results will help determine whether such comparisons are possible or desirable.

While experts do not recommend comparing across languages using some sort of universal standard for, say, correct words per minute by the end of grade 2, approximate comparisons within languages appear to be possible, and in any case, for items such as fluency of letter recognition in languages using a Latin script, basic comparisons should be possible.

III. 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), the National Literacy Panel (2004), and the Committee on the Prevention of Reading Difficulties in Young Children (Snow, Burns, & Griffin, 1998), among others. Each of these works highlights the key components for early reading acquisition and instruction.

Although this brief summary certainly does not do justice to the entire field of reading research, a basic understanding of the research foundations is critical to the understanding of EGRA.

The two main principles derived from this body of literature that support the development of EGRA are as follows. First, reading assessment (and its complement, instruction) is complex, but there is sufficient research evidence to support the development of specific assessment tools to determine what skills students need in order to become successful readers, *regardless of the method by which students are being taught*. Second, early reading skills are acquired in phases; the level of complexity of a language affects how long students need to acquire early reading skills. Each of these principles is explained in detail, below.

Assessing Early Reading

Assessing early reading acquisition is complicated, but we know what skills to focus on. We can derive an understanding of the important aspects of assessment from the critical components of instruction. As Torgesen (1998) states: Adequate reading comprehension is the most important ultimate outcome of effective instruction in reading.

Deriving lessons from an exhaustive review of research, consultations with experts, and public hearings, the members of the National Reading Panel (National Institute of Child Health and Human Development, 2000) highlighted five essential components of effective reading instruction, as follows:

(1) phonemic awareness – instruction designed to teach children the ability to focus on, manipulate, and break apart the sounds (or phonemes) in words;
(2) phonics – instruction designed to help readers understand and apply the knowledge of how letters are linked to sounds (phonemes) to form letter-sound (grapheme-phoneme) correspondences and spelling patterns;
(3) fluency – instruction, primarily through guided oral reading, that reinforces the ability to read orally with speed, accuracy, and proper expression;
(4) vocabulary – instruction, both explicit and implicit, in order to increase both oral and print knowledge of words, a critical component of comprehension and reading; and
(5) comprehension – instruction that teaches students to actively engage with, and derive meaning from, the texts they read.

One useful theoretical model derived from this literature can be described as follows: Comprehension is a function of decoding, linguistic comprehension, and speed. Evidently, this is not a "true" equation but a heuristic model, as the variables do not have the same metric, the nature of the coefficients is not specified and, furthermore, decoding and speed are not entirely separate concepts (as decoding itself is only meaningful when it is automatic). However, the model has its underpinnings in research that documents the independent contribution of these various factors. In even simpler expressions of this heuristic, the "speed" component is left out, as "decoding" is taken to mean essentially automatic and instantaneous (sounding like normal speech) recognition and duplication of print-sound relations, with automaticy/decoding being necessary—but not sufficient—for comprehension. This is derived from Gough's (1996) "simple view of reading," namely that comprehension requires (1) general language comprehension ability and (2) ability to accurately and fluently identify words in print. However, more recent expressions of this model do highlight the importance of speed or fluency as having independent, explanatory power (see Carver, 1998; Catts, Hogan, & Fey, 2003; Chiappe, 2006; Cutting & Scarborough, 2006; Ehri, 1998; Gough & Tunmer, 1986; Hoover & Gough, 1990; Joshi & Aaron, 2000; Share, 1999; Sprenger-Charolles, Siegel, Béchennec, & Serniclaes, 2003).

The research foundations underlying EGRA also support literacy assessment for adults. UNESCO's Institute for Statistics is currently developing the Literacy Assessment and Monitoring Programme (LAMP). According to their website:

"LAMP measures five component skills, considered the building blocks of fluent reading:

1. Alphanumeric perceptual knowledge and familiarity: the ability to recognise the letters of the alphabet and single digit numbers.
2. Word recognition: common words, appearing frequently in print, are expected to be in the listening/speaking vocabulary of an individual who speaks the target language.
3. Decoding and sight recognition: the ability to rapidly produce plausible pronunciations of novel or pseudo words by applying sight-to-sound correspondences of the writing system.
4. Sentence processing: the ability to accurately and rapidly process simple, written sentences and apply language skills for comprehension.
5. Passage reading: the ability to process simple written passages and apply language skills for comprehension with ease."

http://www.uis.unesco.org/ev.php?ID=6412_201&ID2=DO_TOPIC

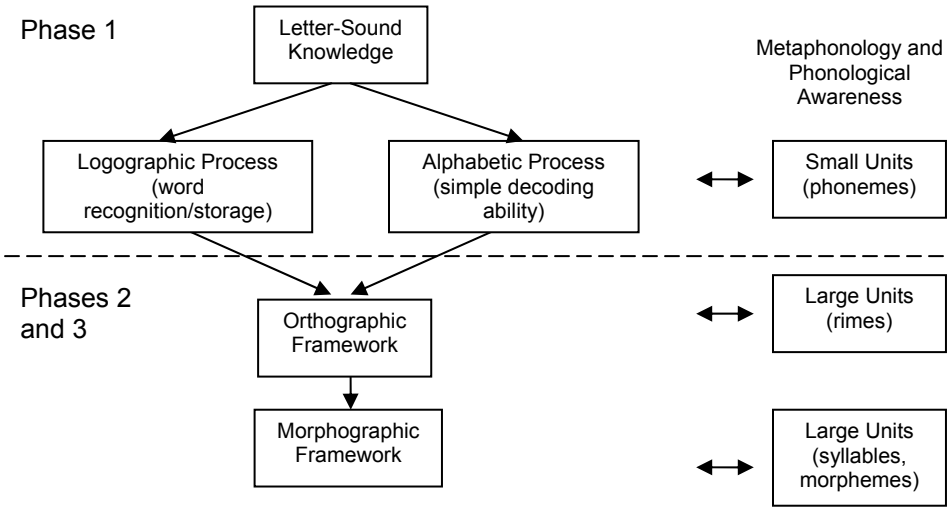
Reading Is Acquired in Phases

A second important generalization from the literature is that acquiring reading skills is a multiphased process that takes a longer time in some languages than others. A schematic of this multiphased learning process is provided in Exhibit 4 below. According to Seymour et al. (2003): "reading is acquired in phases, such that basic foundational components are established in Phase 1, while the complexities of orthographic and morphographic structure are internalized in Phases 2 and 3. The foundation consists of two processes, a logographic process involved in the identification and storage of familiar words, and an alphabetic process which supports sequential decoding" (p. 144).

Research comparing reading acquisition in 13 European languages provides evidence in support of this latter hypothesis: that the regularity of a language impacts the speed of acquisition of foundation reading skills. Level of complexity of a language affects reading acquisition from the very beginning of the learning process. In this case, acquisition of even the basic foundation skills would occur at a slower rate in languages with more complex constructions. That is, it may be the case that the level of complexity of a language affects the rate at which students learn even the most basic of skills, such as letter identification (for reviews of cross-language differences in reading acquisition, see Sprenger-Charolles, Colé, &

Serniclaes, 2006; Zieger & Goswami, 2005). Work resulting from EGRA applications in The Gambia and Senegal seem to support this as well; that is, English performance was much lower than French performance in nearly all of the tested skills.

Exhibit 4. The Dual-Foundation Model of Orthographic Development



Source: Adapted from Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94,* 143-174.[7]

One concrete example for comparing language differences is that of the regularity of both French and English. One of the main problems of the French orthography is not the spelling of vowels (as in English), but the spelling of the ends of words, which are often silent. In English, the consistency of phoneme-grapheme correspondences (PGC, or how sounds correspond to letters or groups of letters) is higher than that of grapheme-phoneme correspondences (GPC, or how letters or groups of letters correspond to sounds). As an example, the /s/ phoneme can be represented by both of the letters "s" and "c" (as in "sit" and "cent") as well as the double-letter grapheme "ss," as in "pass." To make things even more complicated, the letter "s" sounds different in the words "sit," "as," "shin," and "treasure," because the grapheme "s" is associated with at least four different phonemes: /s/ in "sit" /z/ in "as", /ʃ/ in "shin" and /ʒ/ in "treasure" (the symbols are part of the International Phonetic Alphabet; see http://en.wikipedia.org/wiki/Wikipedia:IPA_for_English). Most vowels in English are associated with even more sounds than are consonants: the letter "a" has approximately 11 different possible sounds. There are 26 letters but approximately 44 phonemes in the English language (the number of reported phonemes in English varies depending on the source, country, and region, as different dialects have different pronunciations).

Exhibit 5 details the GPC and PGC correspondences for reading and spelling of vowels in English versus French. That is, in English, "following the rules" for reading and spelling for single-syllable words produces the correct result only 48 percent of the time for reading and 67 percent of the time for spelling. In French, reading demonstrates a high level of regularity of

[7] Note that acquisition may be simultaneous. There is evidence that the developmental process for acquiring phonological/phonemic awareness in English begins with larger phonological units and gradually moves to smaller units. Regardless of the order of phasing, the model highlights the need to acquire all of the foundation skills.

grapheme-phoneme correspondence, but spelling produces nearly as many challenges as does spelling in English (for a review see Sprenger-Charolles et al., 2006).

Exhibit 5. Consistency of Grapheme-Phoneme Correspondences (GPC) and Phoneme-Grapheme Correspondences (PGC) for Vowels in English and French (monosyllabic items)

Language	GPC (reading)	PGC (spelling)
English	48%	67%
French	94%	68%

Source: Peereman, R., & Content, A., (1999). LEXOP: A lexical database providing orthography-phonology statistics for French monosyllabic words. *Behavioral Methods, Instruments and Computers, 31,* 376-379.

That said, English is more predictable than the above would lead us to believe. Knowledge of word origin and other clues can help in the reading process. As referenced in Moats' (2004) *Language essentials for teachers of reading and spelling*:

- 50% of words are predictable by rule:
  cat kit back
  actress segregate struggle

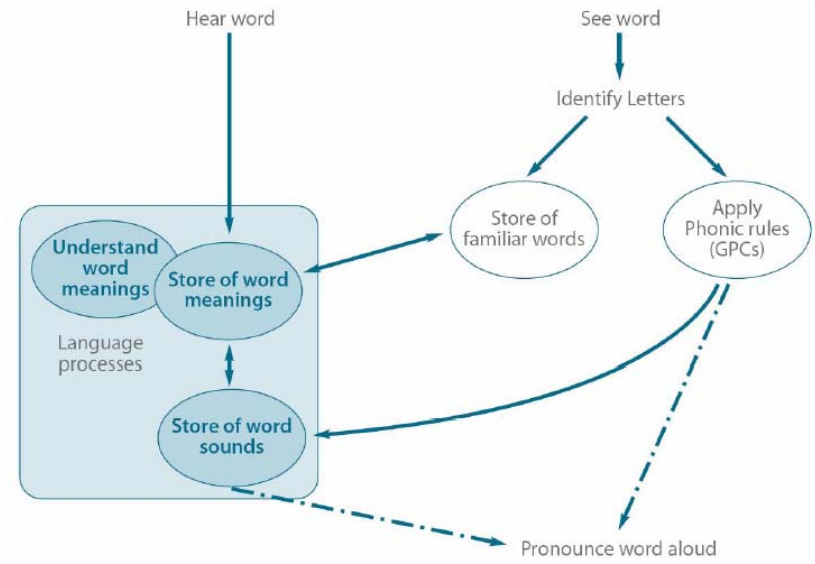
- 36% of words are predictable by rule with one error:
  afraid (*afrade*) friendly (*frendly*) bite (*bight*)
  answer (*anser*) father (*fother*) shoe (*shue*)

- 10% of words will be predictable with morphology and word origin taken into account:
  health/heal anxious/anxiety

- Fewer than 4% are true oddities:
  psyche (*sikee*) physician (*fasishan*)

Finally, the "Rose Report" (Department for Children, Schools and Families, 2006), an independent report examining best practice in teaching reading produced for the United Kingdom's Department of Children, Schools and Families (formerly the Department for Education and Skills), depicts the following process for word recognition (see Exhibit 6 below). When children hear a word, they relate that to their store of word meanings and word sounds, to later process the word and pronounce the word aloud. Developing word recognition skills includes such components as letter recognition, acquisition of essential phonics rules and GPCs, and building of vocabulary, all of which are tested by the EGRA assessment.

Exhibit 6. Word Recognition Process



Note: GPC = grapheme-phoneme correspondence

Source: Department for Children, Schools and Families (2006). *The new conceptual framework for teaching reading: The "simple view of reading."* Overview for literacy readers and managers in schools and early years settings. Retrieved August 2007 from http://www.standards.dfes.gov.uk/eyfs/resources/downloads/paper_on_searchlights_model.pdf

As noted earlier, EGRA's design and emphasis on early reading skills, and in particular phonics and phonemic awareness, means that the initial components of the assessment are not suitable, as designed, for children in upper grades. (This recommendation assumes that students have acquired the foundation skills in the early grades; decisions about use of EGRA should be skill- rather than age-dependent and should rely on local knowledge of students' rate of skill acquisition, if available.) Once the lessons of phonemic awareness and phonics are fully incorporated into the child's reading process, it is no longer appropriate to focus on these skills, through either assessment or instruction. That is, phonics instruction is time-limited, whereas language comprehension and vocabulary instruction are lifelong practices that can and should be both assessed and taught. As children move from "learning to read" to "reading to learn," the balance of instruction will change as well (Espin & Tindal, 1998; International Reading Association, 2007).

IV. EGRA Adaptation and Research Workshop

Once a Ministry or other within-country organization decides to begin the process of adapting and applying the EGRA instrument, the first step is to organize an in-country workshop, normally lasting about five working days, which has as its objectives to

- give both Ministry officials and local curriculum and assessment specialists a grounding in the research backing the instrument components.
- review the informed consent procedures and discuss the ethics of research and working with human subjects, especially children.
- review the instrument components and linkages to instruction.
- adapt the instrument to local conditions using the item construction guidelines provided in this toolkit (including translating the instrument instructions; developing a local-language version, if necessary; and modifying the word and passage reading components to reflect locally and culturally appropriate words and concepts).
- train field supervisors in the supervision of the instrument administration process, including assessing interrater reliability and pretesting the newly adapted instrument.

The section reviews the steps for preparing and delivering an EGRA workshop and provides an overview of the topics to be covered. Following the workshop, a second week of training for enumerators and piloting of the instrument (in teams of four to five per school) is recommended. Enumerator training and fieldwork is discussed in Section V.

Adaptation and Research Workshop

The number of participants in the adaptation and research workshop will be determined by the number of schools to be sampled (see Annex B) and the availability of Ministry staff to participate in both the design and actual fieldwork in schools. The use of Ministry staff is recommended in order to build capacity and help ensure sustainability for these assessments. Participants may also include non-Ministry practitioners; academics; teachers; and experts in curriculum development, assessment, school supervision and support, and local language, if possible. The ideal situation is for Ministry staff to participate throughout the entire adaptation and piloting process (upwards of 1 month in total, depending on the number of schools to be sampled). In countries where Ministry staff members are not available for this length of time, the group should be complemented by contracted field supervisors and enumerators.

Field supervisors should participate in the entire adaptation and research justification process so as to better understand the underlying principles of each of the assessment components. If not all participants are available for the entire adaptation process, additional enumerators may attend a week-long training that follows this workshop and includes piloting in several schools for practice and final modifications to the instrument. A greater number of enumerators than necessary should participate during the follow-up training so as to allow for selection of those that prove to be best able to administer the instrument (determined through calculations of interrater reliability; see Annex C).

Groups composed of Ministry staff, teacher trainers, retired teachers, and teachers in training provide a good mix of experience, enthusiasm, and energy—important elements of the assessment process.

While Ministry staff in particular should be selected based on their ability to contribute to the design and adaptation of the EGRA instruments, it is possible that not all Ministry staff may be selected as enumerators or supervisors. The two most important enumerator qualifications are:

- **The ability to interact in a nonthreatening manner with young children.** As the instrument is oral and individually administered, the quality and accuracy of the data depend largely on the ability of the enumerator to encourage and calm the students such that they perform to the best of their abilities. False-negative results, from students who can perform well but who are nervous or fearful, can be minimized with the right approach. While some of this can be practiced and trained, a large part is attitude and personality, both of which are difficult to modify within a 1- to 2-week training exercise.

- **Organizational skills.** The second important qualification is the ability to handle several tasks at one time, including listening to the student, scoring the results, and operating a stopwatch or timer.

As part of the selection process, workshop leaders should conduct tests of interrater reliability using audio recordings of students participating in pilot assessments. Enumerators listen and score the assessment, then report back their scores. Those enumerators whose scores are more than one standard deviation from the mean should be provided additional training opportunities and practice. If they do not improve they should not be selected for participation in the data collection process. Additional tests of interrater reliability are discussed in Section V.

The workshop should be delivered by a team of at least two experts. The first expert—responsible for leading the adaptation of the instrument, presenting the development process of the EGRA instrument, and guiding the data collection and entry processes—should have a background in education survey research and assessment/test design. This experience should include basic statistics and a working knowledge of Excel and a statistical program such as SPSS or Stata. The second expert—responsible for presenting reading research and pedagogical/instruction processes—should have a background in reading assessment tools and instruction. Both workshop leaders should be well versed in the components and justifications of the assessment and be adept at working in a variety of countries and contexts.

When possible, logistical support (venue, per diems, school visits, printing, laminating of student instruments, etc.) should be provided by a local NGO or firm accustomed to conducting surveys in local schools.

Materials for the workshop include:

- Paper and pencils with erasers for participants
- Stopwatches or timers (if possible, find a kitchen timer that counts down from one minute)[8]
- LCD projector, whiteboard, and flipchart (if possible, the LCD projector should be able to project onto the whiteboard for simulated scoring exercises)
- Copies of the presentations, supervisor manual, and draft instruments
- Presentation on the EGRA development process, purpose, uses, and research background
- Presentation on reading instruction research related to EGRA.

A sample agenda for the adaptation and research workshop is presented in Exhibit 7.

Exhibit 7. Sample Agenda: EGRA Adaptation and Research Workshop

Monday	Tuesday	Wednesday	Thursday	Friday
• Introduce facilitators and attendees • Review logistics • Review EGRA development process • Review research underlying EGRA	• Review draft instrument and student questionnaire • Revise draft instrument	• Train in use of application, and practice	• Pretest in 2-6 schools • Enter data as a practice exercise • Test for inter-rater reliability • Conduct simple analyses using Excel	• Discuss results • Discuss instructional implications • Revise draft instrument based on results

Note on Instrument Design and Coding Strategy

There is one overarching concern regarding the instrument components and other inputs about the design that ultimately will have a significant effect on the data coding and analysis. The concern is that the coding system for the items in the instrument must allow evaluators to differentiate among several types of similar-looking responses. For example, there may be situations in which (a) a question was not asked, (b) a student did not know the answer, (c) a student would not answer or could not answer, or (d) the student's answer was completely incorrect (a true zero). A common mistake is to fail to distinguish among all these. No universally valid set of values is mandated for these responses, however, so decisions like these have to be tailored to the instrument under development. The current versions of the EGRA instrument reflect this coding strategy, including stop rules and no-response codes.

Note on Ethics of Research and Institutional Review Board (IRB)

As a research institution receiving federal grants, RTI follows the U.S. federal regulations for conducting ethical research. As noted in RTI's description of the process: Institutional Review Boards (IRBs) must be utilized by all organizations that conduct research involving human subjects. IRBs use the set of basic principles outlined in the "Belmont Report," a report issued in 1978 by the United States National Commission for the Protection of Human Subjects of

[8] Although somewhat expensive, the best stopwatch we have found for this purpose is available from the Dynamic Indicators of Basic Early Literacy (DIBELS) website: http://dibels.com/merch.html. Another slightly less costly option is available at http://www.cutleryandmore.com/details.asp?SKU=7499.

Biomedical and Behavioral Research, 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.
- **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. (RTI internal website, accessed January 12, 2009)

For each of the assessments conducted to date, RTI has included a verbal consent for human subjects participating in the assessments. Prior to administering the assessment, enumerators 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. If school principal or teacher surveys are conducted as part of the study, a similar written consent process is completed. While this consent process is often unfamiliar to local country counterparts, the process is often welcomed by students and teachers who report feeling empowered at being given the option to participate in the assessment. Few students and teachers decline to participate (in the case of a recent sample from Nicaragua of more than 6,000 students, only 8 declined to participate). In these cases another student is randomly selected. For additional information on IRBs and ethical research with human subjects, including children, please see http://www.hhs.gov/ohrp/faq.html.

Review of the Instrument Components

To develop the complete Early Grade Reading Assessment, the EGRA development team reviewed more than a dozen assessment instruments, including the Dynamic Indicators of Basic Early Literacy (DIBELS), the Peabody Picture Vocabulary Test, and instruments applied in Spain, Peru, Kenya, Mongolia, and India. The EGRA instrument also builds on lessons from ongoing efforts to develop “smaller, quicker, cheaper” ways to assess (adult) literacy (International Literacy Institute & UNESCO, 2002; Wagner, 2003).

As discussed above, to obtain feedback on the initial design of EGRA, USAID, the World Bank, and RTI hosted a meeting of experts (a summary of proceedings and a list of workshop participants can be found at www.eddataglobal.org, under News and Events). Based on this and other expert consultations, a complete Early Grade Reading Assessment was developed for application in English. The resulting instrument contains eight tasks, or subtests, as follows:

1. Letter name knowledge
2. Phonemic awareness
3. Letter sound knowledge
4. Familiar word reading
5. Unfamiliar word reading
6. Oral reading fluency with comprehension
7. Listening comprehension
8. Dictation.

Each of these components has been piloted in Arabic, English, French, Spanish, and several other languages through both World Bank and USAID-funded initiatives. 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 foundation reading skills to be able to identify which areas need additional instruction. If EGRA only tested oral fluency, many low-income country results would have considerable problems with 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). For example, the average third-grade student tested in one African country was below the 10th percentile for a first-grade student in the United States (both tested at the end of the school year). This indicates that only testing oral reading fluency would not yield sufficient information to inform Ministry staff about what is or is not going on in terms of which pre-reading skills need improvement (letter recognition, etc.).

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. 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 to the procedures, be justified and documented. While RTI, the World Bank, and USAID understand that different donors and countries will adapt the instrument to their own needs, it is important that such changes be justified and explained in terms of the purpose and use of the assessment.[9]

For a detailed explanation of the technical quality and reliability of the EGRA instrument, including guidelines for conducting basic instrument quality and reliability checks, please see Annex C of this toolkit.

To summarize the overall EGRA assessment approach to workshop participants, Exhibit 8 should be shared during the review of each of the individual instrument components. Participants should also understand the difference between (1) testing of student abilities and (2) instructional techniques for improving student performance in these skills. That is, as discussed above, students should not be taught the test components; rather, instructional approaches based on the EGRA results should be developed. Additional discussion of instructional approaches can be found in Section VII.

Exhibit 8. Review of Instrument Components

Component	Early reading skill	Skill demonstrated by students' ability to:
1. Letter name knowledge	Letter recognition	• Provide the name of upper- and lowercase letters in random order
2. Phonemic awareness	Phonemic awareness	• Segment words into phonemes • Identify the initial sounds in different words
3. Letter sound knowledge	Phonics	• Provide the sound of upper- and lowercase letters distributed in random order
4. Familiar word reading	Word reading	• Read simple and common one- and two-syllable words
5. Unfamiliar nonword reading	Alphabetic principle	• Make grapheme-phoneme correspondences (GPCs) through the reading of simple nonsense words

[9] RTI and its funders also request that new iterations of the instruments be shared via the EdData II website (www.eddataglobal.org, under the EGRA link) such that the entire education community can learn from the results. All of the instruments developed by RTI to date are freely available on the website; it is RTI's expectation that others will be interested in sharing their instruments and learning processes as well.

Component	Early reading skill	Skill demonstrated by students' ability to:
6. 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
7. Listening comprehension	Listening comprehension	• Respond correctly to different types of questions including literal and inferential questions about the text the enumerator reads to them
8. Dictation	Alphabetic principle	• Write, spell, and use grammar properly through a dictation exercise

1. Letter Name Knowledge

The test of letter name knowledge is the most basic of assessments of student reading preparedness (and risk). Letter name knowledge is a consistent predictor of reading development for native speakers of English, French, and other alphabetic languages (Chiappe, Siegel, & Wade-Woolley, 2002). It has also proved to be a useful indicator for nonnative speakers (Chiappe, 2006).

In this assessment of letter name knowledge, students are asked to provide the names (not the sounds) of all of the letters that they can, within a one-minute period. 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 most similar to standard children's textbooks) in horizontal rows with each letter presented multiple times. Letters are to be selected based on the frequency with which the letter occurs in the language in question (see frequency table below, Exhibit 9). Randomization is used to prevent students from reciting a memorized alphabet—that is, to test for actual automaticity of letter recognition and translation of print to sound. The complete alphabet (both upper- and lowercase) is presented based on evidence that student reading skills in European languages advanced only after about 80 percent of the alphabet was known (Seymour et al., 2003).

Exhibit 9. Letters in English Language: Frequency of Use

E	11.1607%	C	4.5388%	Y	1.7779%
A	8.4966%	U	3.6308%	W	1.2899%
R	7.5809%	D	3.3844%	K	1.1016%
I	7.5448%	P	3.1671%	V	1.0074%
O	7.1635%	M	3.0129%	X	0.2902%
T	6.9509%	H	3.0034%	Z	0.2722%
N	6.6544%	G	2.4705%	J	0.1965%
S	5.7351%	B	2.0720%	Q	0.1962%
L	5.4893%	F	1.8121%		

Source: AskOxford.com. (n.d.) *Ask the experts: Frequently asked questions, Words.* Retrieved July 2008, from http://www.askoxford.com/asktheexperts/faq/aboutwords/frequency?view=uk

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 common alphabetic languages.[10] Test developers constructing instruments in local languages should analyze electronic texts to develop similar letter frequency tables. To develop a letter frequency table, take a large representative document in Microsoft Word and use the "Find" command (under the Edit menu in Office 2003; or Ctrl+F). Enter the letter "a" in the "Find what" box and check the box to "Highlight all items found" in the main document. Click on "Find all" and Microsoft Word 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 toolkit, for example, the letter "a" appears nearly 14,000 times). Repeat this process for each letter of the alphabet, recording the total number for each letter until you can calculate the proportion each letter appears as a share of the total number of letters in the document.

Pronunciation issues need to be handled with sensitivity in this and other tasks. The issue here is not to test for "correct" pronunciation, where "correctness" is interpreted as hewing to some standard that indicates privileged socioeconomic status. The issue is to test automaticity using a pronunciation that may be common in a given region or form of English. Thus, regional accents are acceptable in judging whether a letter is named correctly.

Data. The child's score for this subtest should be calculated as the number of correct letters per minute. If the child completes all of the words before the time expires, the time of completion should be recorded and the calculations should be based on that time period. Enumerators should mark any incorrect letters with a slash (***/***), place a bracket (**]**) after the last letter named, and record the time remaining on the stopwatch at the completion of the exercise (variables are thus: Total letters read, Total incorrect, Time remaining on stopwatch). These three data points are then used to calculate the total correct letters per minute (CLPM):

CLPM = (Total letters read – Total incorrect) / [(60 – Time remaining on stopwatch) / 60]

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

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

In addition, for each of the timed tasks, below, enumerators should only record the information noted above. Having enumerators calculate results such as CLPM in the field distracts from the evaluation process and can lead to significant errors.

Item Construction. Letters of the alphabet should be distributed randomly, 10 to a line and should be evenly distributed among upper- and lowercase letters. The percentages should act as a guide for the frequency with which the letters appear in the task sheet (i.e., for English, in a list of 100 letters, the letter "E" should appear approximately 11 times, the letter "A" 8 times, etc.).

[10] Spanish, French, German, Italian, Esperanto, Turkish, and Swedish are available at http://en.wikipedia.org/wiki/Letter_frequencies#Relative_frequencies_of_letters_in_other_languages (accessed February 10, 2009).

Sample Assessment Design: Letter Name Knowledge

Show the child the sheet of letters in the student stimuli booklet. Say:

Here is a page full of letters of the alphabet. Please tell me the NAMES of as many letters as you can--not the SOUNDS of the letters, but the names.

For example, the name of this letter *[point to A]* **is "A"**

Let's practise: tell me the name of this letter *[point to V]:*
If the child responds correctly say: **Good, the name of this letter is "VEE."**
If the child does not respond correctly, say: **The name of this letter is "VEE."**
Now try another one: tell me the name of this letter *[point to L]:*
If the child responds correctly say: **Good, the name of this letter is "ELL."**
If the child does not respond correctly, say: **The name of this letter is "ELL."**

Do you understand what you are to do?

When I say "Begin," please name the letters as quickly and carefully as you can. Start here and continue this way. *[Point to the first letter on the row after the example and draw your finger across the first line].* **If you come to a letter you do not know, I will tell it to you. Otherwise I will keep quiet & listen to you. Ready? Begin.**

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've already marked the self-corrected letter as incorrect, circle the letter and go on.* ***Stay quiet,*** *except when providing answers as follows: if the child hesitates for 3 seconds, provide the name of the letter, point to the next letter and say "****Please go on****." Mark the letter you provide to the child 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 exercise.*

AFTER 60 SECONDS SAY, "stop." Mark the final letter read with a bracket (⌋).
Early stop rule: *If the child does not give a single correct response on the first line, say "****Thank you!****", discontinue this exercise, check the box at the bottom, and go on to the next exercise.*

Example : A v L

1	2	3	4	5	6	7	8	9	10	
L	i	h	R	S	y	E	O	n	T	*(10)*
i	e	T	D	A	t	a	d	e	w	*(20)*
h	O	e	m	U	r	L	G	R	u	*(30)*
g	R	B	E	i	f	m	t	s	r	*(40)*
S	T	C	N	p	A	F	c	a	E	*(50)*
y	s	Q	A	M	C	O	t	n	P	*(60)*
e	A	e	s	O	F	h	u	A	t	*(70)*
R	q	H	b	S	i	g	m	i	L	*(80)*
L	i	N	O	e	o	E	r	p	X	*(90)*
N	A	c	D	d	I	O	j	e	n	*(100)*

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

Check this box if the exercise was discontinued because the child had no correct answers in the first line. ☐

2. Phonemic Awareness

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. This ability to identify sounds in words, to separate words into sounds, and to manipulate those sounds is termed phonemic awareness, and is a subset of phonological awareness, a more general appreciation of the sounds of speech as distinct from their meaning (Snow et al., 1998). 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." Research has found that phonemic awareness plays an important role in reading acquisition. It has been shown to be the number one predictor of success in reading, better than socioeconomic status, preschool attendance, or reading time in the home (Share, Jorm, Maclearn, & Matthews, 1984).Testing for and remediating this skill is thus important for later reading development.

Thus far, EGRA has piloted an assessment of phonemic awareness in two different ways: using phoneme segmentation and identification of onset and rime sounds (first and last sounds). Each of these approaches, described below, is common in tests of early reading, including:

- DIBELS www.dibels.uoregon.edu
- Test of Phonological Awareness (TOPA) http://www.linguisystems.com/itemdetail.php?id=658
- Comprehensive Test of Phonological Processing (CTOPP) http://ags.pearsonassessments.com/group.asp?nGroupInfoID=a9660

Phoneme segmentation, in this case 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 one of the most predictive of later learning skills. Thus far, phoneme segmentation has proved difficult to administer and has demonstrated large floor-effect problems. It has been included in the toolkit as an example for testing during the pilot phase—that is, if floor-effect problems arise, then a more simple task, such as initial sound identification, should be conducted. Identification of initial sounds has been piloted in English in Guyana and Liberia.

First Approach: Phoneme Segmentation. For this portion of the assessment, the examiner reads aloud a list of 10 simple, one-syllable words, one at a time. Students are asked to identify and sound out each sound present in the word (as this is an auditory assessment there is no student handout, only an examiner coded sheet).

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 two-, three-, and four-phoneme words should be selected. Words should use reasonably common phoneme constructions (minimize the number of complex graphemes—i.e., with more than one letter—and blends).

Sample Assessment Design: Phonemic Awareness

This is NOT a timed exercise and ***THERE IS NO STUDENT SHEET****. Read aloud each word twice, and have the student say the sounds. Remember to model the "pure" sounds: /p/, not "puh" or "pay." Say:*

This is a listening exercise. You know that each letter has a sound. For example, "pot", "p"-"o"-"t" can be sounded out " /p/ - /o/ - /t/ ". I will say a word two times. Listen to the word, then tell me the sounds in that 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 5 seconds for the student to respond].*

Now 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/ /ss/.**
[If the child does not respond correctly, say]: **The sounds in "miss" are: /m/ /i/ /ss/. Now it's your turn. Tell me the sounds in "miss".** *[Wait 5 seconds for the student to respond].*

All right, let's go. I will say a word two times. Listen to the word, then tell me the sounds in that word. Do you understand what you are to do?

Pronounce each word ***TWICE****, slowly (about 1 word per second).*

Put a slash (/) through each incorrect sound as well as any sounds that the child skips.

If the child does not respond to a word after 5 seconds, mark all sounds in the word as incorrect and move on.

Early stop rule: *If the child gives no correct answer among the first* ***five*** *words, say "**Thank you!**", discontinue this exercise, check the box at the bottom of the page, and move on to the next exercise.*

What are the sounds in _____? _____! *[Repeat the word twice]*

"too"	/t/	/oo/		
"say"	/s/	/a/		
"up"	/u/	/p/		
"buy"	/b/	/i/		
"me"	/m/	/ee/		*(5 words)*
"set"	/s/	/e/	/t/	
"mop"	/m/	/o/	/p/	
"jam"	/j/	/a/	/m/	
"fish"	/f/	/i/	/sh/	
"cool"	/k/	/oo/	/l/	

Check this box if the exercise was discontinued because the child had no correct response on the first 5 words. ☐

.

Second Approach: Initial Sound Identification. A second approach to assessing phonemic awareness is to have students identify the first sound in a selection of common words. The example below uses 10 sets of simple words and asks students to identify the initial sound in each of the words. The enumerator 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.

Item Construction. Simple, one-syllable words should be selected from first- or second-grade word lists. It is recommended to use only one-syllable words so as not to overtax students' working memory.

Sample Assessment Design: Initial Sound Identification

This is NOT a timed exercise and ***THERE IS NO STUDENT SHEET****. Read aloud each word twice, and have the student say the sounds. Remember to model the "pure" sounds: /p/, not "puh" or "pay." Say:*

This is a listening exercise. I want you to tell me the beginning sound of each word. For example, in the word "pot", the first sound is "/p/". In this exercise, 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 practise. What is the first sound in "mouse"? "Mouse."
[If the child responds correctly, say]: **Very good, the first sound in "mouse" is /mmmmm/.**
[If the child does not respond correctly, say]: **Listen again: "mmmouse". The first sound in "mouse" is /mmmmm/."**

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 / ".**

Do you understand what you are to do?

Read the prompt and then pronounce the target word a second time. Accept only as correct the isolated sound (without a shwah). If the child does not respond after 3 seconds, mark as "No response" and say the next prompt. Enunciate clearly, but do not overemphasize the beginning sound of each word.

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

What is the first sound in "_____"? "_____"? *[Repeat the word twice]*					
map	/mmmm/	○ Correct	○ Incorrect	○ No Response	
say	/ssssss/	○ Correct	○ Incorrect	○ No Response	
up	/uh/	○ Correct	○ Incorrect	○ No Response	
go	/g'/	○ Correct	○ Incorrect	○ No Response	
now	/nnnn/	○ Correct	○ Incorrect	○ No Response	*(5 words)*

can	/k'/	○ Correct	○ Incorrect	○ No Response
fish	/ffffff/	○ Correct	○ Incorrect	○ No Response
pot	/p'/	○ Correct	○ Incorrect	○ No Response
run	/rrrrrr/	○ Correct	○ Incorrect	○ No Response
look	/llllll/	○ Correct	○ Incorrect	○ No Response

Check this box if the exercise was discontinued because the child had no correct answers in the first five words : ☐

3. Letter sound knowledge

Knowledge of how letters correspond to sounds is another critical skill children must master to become successful readers. Letter-sound correspondences are typically taught through phonics-based approaches, which have moved in and out of favor in the past several decades. This subtest, like the unfamiliar nonword decoding exercise, is likely to be somewhat controversial among some groups of educators. Letter sound 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 covers a range of letters and graphemes, including single consonants and vowels as well as vowel digraphs and dipthongs (i.e., ea, ai, ow, oy).

Data. As in the letter naming exercise, the child's score for this subtest should be calculated based on the number of correct letter sounds per minute.

Item Construction. The same laminated page of letters used in the first subtest of letter name knowledge should be used for assessing letter sound knowledge. For consonants that can represent more than one sound (i.e., c, g), either answer is acceptable. For vowels, either the short or long sound is accepted (/i/ as in pin or as in pine). Students may have difficulty in eliminating the vowel sound frequently associated with consonants; in these cases either /b/ or /buh/ is accepted as a correct response. During training, enumerators and supervisors should carefully review possible pronunciations of each letter. (For a complete listing of characters and symbols in phonetic alphabets, please see http://www.antimoon.com/misc/phonchart2008.pdf.)

4. Familiar Word Reading

Children's decoding 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 comprehension paragraphs, as children are unable to guess the next word from the context. For this assessment, familiar words should be high-frequency words selected from early grade reading materials and storybooks for first-, second-, and third-grade materials (progressively increasing in difficulty). Sources for such word lists abound. In English, Zeno, Ivenz, Millard and Duvvuri's work (1995) is based on a corpus of 17 million English words.

Data. The enumerator records the number of correct words per minute. If the child completes all of the words before time expires, the time of completion should be recorded and the calculations should be based on that time period. Correct words per minute should be recorded and scored.

The same three variables collected for the letter naming exercise, above, should be collected for this and the other timed exercises, namely: Total words read, Total incorrect words, Time remaining on stopwatch. See above discussion for calculations.

Item Construction. Word lists, if not available in country, can be found online (e.g., http://www.english-zone.com/reading/dolch.html). The Dolch English word list includes 220 of the most frequently used words in children's books in the United States (Dolch, 1948). Also called "sight" words (words that primary school children should recognize on sight, as many of these words are not easy to sound out and thus must be memorized), these word lists usually include regular one- and two-syllable words (Moats, 2000). Words should be arranged horizontally with good separation and clear, familiar (lowercase) print in 10 rows, five words per line. The font used should be 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 purchased.

Sample Assessment Design: Familiar Word Identification

Show the child the sheet of familiar words in the student stimuli booklet. Say,

Here are some words. I would like you to read to me as many words as you can (do not spell the words, but read them). For example, this word is: "cat".

Let's practise: 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," read the words as quickly and carefully as you can. Read the words across the page, starting at the first row below the line. I will keep quiet and listen to you, unless you need help. Do you understand what you are to do? Ready? Begin.

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've already marked the self-corrected letter as incorrect, circle the letter and go on. ***Stay quiet,*** *except when providing answers as follows: if the child hesitates for 3 seconds, provide the word, point to the next word and say "****Please go on****." Mark the word you provide to the child as incorrect.*

AFTER 60 SECONDS, SAY "stop." Mark the final word read with a bracket ( ]).
Early stop rule: *If you have slashed/marked as incorrect all of the answers on the first line, say "****Thank you!****", discontinue this exercise, check the box at the bottom, and go on to the next exercise.*

Example : cat sick made

1	2	3	4	5	
go	sad	up	find	come	*(5)*
help	two	run	see	down	*(10)*
red	and	play	at	you	*(15)*
chair	man	when	now	under	*(20)*
please	soon	like	they	good	*(25)*
thank	going	are	know	him	*(30)*
jump	once	ask	fly	want	*(35)*
must	green	sing	those	always	*(40)*
many	which	upon	sit	clean	*(45)*
stop	big	me	house	girl	*(50)*

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

Check this box if the exercise was discontinued because the child had no correct answers in the first line. ☐

5. Unfamiliar Nonword Reading

Pseudoword or nonword reading is a measure of decoding ability and is designed to avoid the problem of sight recognition of words. 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 sight recognition skills; tests that do not include a decoding exercise can overestimate children's ability to read unfamiliar words (as the words tested may be part of the sight recognition vocabulary).

Data. The child's score is calculated as the number of correct nonwords per minute. If the child completes all of the words before time expires, the time of completion should be recorded and the calculations should be based on that time period. The same three variables collected for the letter naming and word reading exercise, above, should be collected for this and the other timed exercises, namely: Total nonwords read, Total incorrect nonwords, Time remaining on stopwatch. See above discussion for calculations.

Item Construction. This portion of the assessment should include a list of 50 one- and two-syllable nonwords, five per row, with the following patterns of letters (C = consonant, V = vowel): CV, VC, CVC. (This may be adjustable by language.) Forms should be legal for the language, using letters in legitimate positions (e.g., not "wuj" because "j" is not used as a final letter in English), should stick to consonant-vowel combinations that are typical of the language, and

should not be homophones of real words (not “kab,” homophone of “cab”). They should be arranged in rows (five nonwords per row), using clear, well-spaced print.

Sample Assessment Design: Unfamiliar Nonword Decoding

Show the child the sheet of invented words in the student stimuli booklet. Say,

Here are some made-up words. 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 next word:* **dif** *].*
[If the student says “dif”, say]: **“Very good: “dif”**
[If the student does not say “dif” correctly say]: **This made-up word is “dif.”**

Now try another one: please read this word *[point to the next word:* **mab** *].*
[If the student says “mab”, say]: **“Very good: “mab”**
[If the student does not say “mab” correctly say]: **This made-up word is “mab.”**

When I say “begin,” read the words as quickly and carefully as you can. Read the words across the page, starting at the first row below the line. I will keep quiet and listen to you, unless you need help. Do you understand what you are to do? Ready? Begin.

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’ve already marked the self-corrected letter as incorrect, circle the letter and go on. ***Stay quiet,*** *except when providing answers as follows: if the child hesitates for 3 seconds, provide the word, point to the next word and say “****Please go on****.” Mark the word you provide to the child as incorrect.*

AFTER 60 SECONDS, SAY “Stop.” Mark the final word read with a bracket ( ]).

Early stop rule: *If you have slashed/marked as incorrect all of the answers on the first line, say “****Thank you!****”, discontinue this exercise, check the box at the bottom, and go on to the next exercise.*

Example : ut dif mab

1	*2*	*3*	*4*	*5*	
fut	lus	dit	leb	gak	*(5)*
huz	jod	kib	lek	tob	*(10)*
nom	rop	hig	reg	san	*(15)*
tup	ral	wix	nep	nad	*(20)*
lut	yod	sim	tat	sig	*(25)*
en	mon	nup	sen	kad	*(30)*
taw	lew	paf	sal	zuv	*(35)*
ved	kag	vom	riz	gof	*(40)*
maz	kol	ver	et	beb	*(45)*
tib	lef	yag	lim	dov	*(50)*

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

Check this box if the exercise was discontinued because the child had no correct answers in the first line. ☐

6. Passage Reading and 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 skill. 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). Poor performance on a reading comprehension tool would suggest that the student had 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 three student scores: the proportion of words read, time per word, and proportion of questions correctly answered. The same three variables collected for the letter naming, word reading, and nonsense word reading exercises, above, should be collected for this and the other timed exercises, namely: Total words read, Total incorrect words, Time remaining on stopwatch. See above discussion for calculations. In addition, results for each of the comprehension questions should be collected and entered into the database, with a final score variable calculated as a share of total questions asked. Questions should only be asked for the text the child has read (see structure of questions and paragraph below).

Item Construction. To create the assessment, examiners should review one-paragraph narratives from children's reading materials (not the school textbook). A narrative story should have 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 should not be a list of loosely connected sentences. Typical character names from the school textbook should be avoided as students may give automated responses based on the stories with which they are familiar. Names and places should reflect the local culture and narratives should have a main character, beginning, middle, and end. Texts should contain some complex vocabulary (inflected forms, derivations, etc.) and sentence structures. Large, clear, familiar print and good spacing between lines should be used to facilitate student reading. No pictures should be included. Comprehension questions should include choice and fact-based questions as well as at least one question requiring inference from the text.

Sample Assessment Design: Passage Reading and Comprehension

Show the child the story on the last page of the student form. Say,

Here is a short story. I want you to read it aloud. When you have finished, I will ask you some questions about what you have read. Do you understand what you are to do? When I say "begin," read the story as quickly and carefully as you can. I will keep quiet and listen to you, unless you need help. Ready? Begin.

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. ***Stay quiet,*** *unless the child hesitates for 3 seconds, in which case provide the word, point to the next word and say "****Please go on****." Mark the word you provide to the child as incorrect.*

At 60 seconds, say "Stop." Mark the final word read with a bracket ( ]).

Early stop rule: *If the child gives no correct answers on the first line, say "****Thank you!****", discontinue this exercise, check the box at the bottom of the page, and go on to the next exercise.*

When 60 seconds are up or if the child finishes reading the passage in less than 60 seconds, ***REMOVE the passage from in front of the child****, and ask the first question below. Give the child at most 15 seconds to answer the question, mark the child's response, and move to the next question. Read the questions for each line up to the bracket showing where the child stopped reading.*

		Now I am going to ask you a few questions about the story you just read. Try to answer the questions as best as you can.	Correct	Incorrect	No Response
My name is Pat. I live on a farm with my mother, father and brother Sam.	16	How many people are in Pat's family? [four; mother, father, Sam, Pat]			
My family is happy. One day Sam planted seeds with father. A snake bit him!	31	Sam's family was happy, then something happened. What was it? [a snake bit Sam]			
Mother knew what to do. She put wet cloths and leaves on his leg.	45	Where did the snake bite Sam? [leg] How did Sam get well? [his mother helped him; put wet cloths and leaves on his leg			
The next day, Sam was back in the field with father. We were happy again.	60	Was made the family happy again? [Sam was healed] What do you think happened to the snake? [ran off; Pat or father killed it; or any reasonable answer.]			

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

Check this box if the exercise was discontinued because the child had no correct answers in the first line. ☐

7. Listening Comprehension

A listening comprehension assessment involves passages that are read aloud by the enumerator; students then respond to oral comprehension questions or statements. Testing of listening comprehension separately from reading comprehension is important due to the different ways in which learners approach, process, and respond to text. 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). The purpose of this assessment is to see whether the student can listen to a passage being read and then answer several questions correctly with a word or a simple statement. Poor performance on a listening comprehension tool would suggest that children simply do not have the basic vocabulary that the reading materials expect, or that they have difficulty processing what they hear.

Data. Students are scored on the number of correct statements they give as their answers (out of the total number of questions). Instrument designers should avoid questions with only "yes" or "no" answers.

Item Construction. Passages should be about 30 words in length and narrate an activity or event that will be familiar to local children. Choice and inference questions should be included.

Sample Assessment Design: Listening Comprehension

This is NOT a timed exercise and ***THERE IS NO STUDENT SHEET.*** *Read the following passage aloud to the child* ***ONLY ONE TIME****, slowly (about 1 word per second). Say,*

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. Do you understand what you are to do?

The gray duckling fell in the mud. "Help me," she cried. A green frog came to help, but he fell in too. "What now?" asked the frog. "I see something that leads to land!" the duckling replied. They both climbed on the log. "We are saved!" they shouted.

What is a duckling?	[duck; little duck]	○ **Correct**	○ **Incorrect**	○ **No Response**
Who fell in the mud last?	[frog]	○ **Correct**	○ **Incorrect**	○ **No Response**
What did the duckling see that was important?	[log]	○ **Correct**	○ **Incorrect**	○ **No Response**
How did the frog and the duck get out of the mud?	[climbed on log]	○ **Correct**	○ **Incorrect**	○ **No Response**
Why do you think the frog and the duck are friends?	[tried to help; helped each other]	○ **Correct**	○ **Incorrect**	○ **No Response**

8. Dictation

Dictation assessment is frequently used by teachers to test both oral comprehension and writing skills. As discussed above, 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 assessment is inspired by models promoted by the Educational Testing Service (2005) and Children's Literacy Initiative (2000) and supported by research by the International Reading Association (Denton, Ciancio, & Fletcher, 2006).

Data. Students are scored on a simple scale that captures accuracy for vowel and consonant sounds, spelling, spacing and direction of text, capitalization, and punctuation. Each category has a total of 2 possible points for total accuracy, with 1 for some accuracy and 0 for no accuracy (see scoring rubric in the sample assessment design, below). During analysis, these variables are added up for a single score variable.

Item Construction. The dictation sentence should be at most 10 words in length and contain at least one difficult or irregular word.

Sample Assessment Design: Dictation

Turn this student response form to the last, lined page for writing, and place it in front of the student. Take the student stimulus sheet and turn to the last page, where you will find the same instructions as below. Say,

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 then read it again so that you can check your work. Do you understand what you are to do?

The student will write the dictation sentence on the lined page of the response form. ***Read the following sentence aloud ONCE at about 1 word per second. Then give the child a pencil, and repeat a SECOND time, grouping the words "Go to the shop" - "and buy some rice" - "and sugar". Pause for five seconds then repeat the sentence a THIRD time while the child is writing.*** *Give the child up to 15 seconds to complete writing after the third reading. Thank the child for participating.*

Go to the shop and buy some rice and sugar.

CODING FOR DATA ENTRY PERSONNEL ONLY—DO NOT CODE AT THE SCHOOL.				
Evaluation Criteria	**Score**	**Correct = 2**	**Partially Correct = 1**	**Incorrect = 0**
Wrote "shop" correctly.			*(sh, ho, op, sho, sop)*	
Wrote "buy" correctly.			*(bu, uy)*	
Wrote "rice" correctly.			*(ri, ic, ce, ric, ice)*	
Wrote "sugar" correctly.			*(sug, uga, gar, suga, ugar)*	
Used spacing between words (size of spacing does not matter).		*9 spaces (between all words)*	*5-8 spaces in the sentence.*	*0-4 spaces.*
Used appropriate direction of text (left to right).			DO NOT MARK HERE	
Used capital letter for the word "Go"			DO NOT MARK HERE	
Used full stop (.) at end of sentence.			DO NOT MARK HERE	

Other Potential Instrument Components and Reasons for Exclusion

During 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 submeasure in predicting future student success in reading. As there is little literature in the developing world (and across multiple languages) regarding the application of these measures, RTI relied on the existing literature, which is mainly from the United States and Europe, although some literature is available for Latin America.

Foremost among these suggestions was the inclusion of a picture-based subtest 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 that of body parts: hand, head, toe, etc.). At this stage, EGRA does not include pictures or picture vocabulary tests for several reasons: (1) Vocabulary is indirectly measured in both the word 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) the difficulty in creating pictures that are universally appropriate for all cultures and contexts was an important consideration. In addition, when pictures are locally developed and crafted, in order to avoid copyright problems or to make them culturally appropriate, at least two problems seem to arise. First, 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.

One of the components initially tested and later eliminated from the assessment was derived from Marie Clay's (1993) Concepts About Print assessment. In The Gambia, Senegal, and Nicaragua, use of three of Clay's items (3 through 5, directional rules including 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* reports 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 segment, either in a reduced or full form (the full battery contains 24 items).

Translation

The consensus emerging among experts such as those convened at the November 2006 Washington meeting, as well as Penny Chiappe at a design session with the South African Department of Education, is that when evaluators are looking for ways to use EGRA in home languages, it is not a good idea to simply translate either the words or the connected-text passage from a base English version (or any other language version) into the home language. Translation may result in very long words in a home language, for instance. Instead, the recommendation is that a passage of approximately equal difficulty to the base English (or Spanish or French, depending on the country in question) passage be used. Simple one- and two-syllable words, short sentences, and a familiar narrative should be used. To the degree that the reading texts have been validated for correspondence with the national curriculum, using passages from texts will also tend to help skirt issues of validity or appropriateness of choice of reading passages. An alternative is to ask teachers and curriculum experts versed in the rules of the home languages to craft a passage that is similar in level of difficulty to the English passage.

As noted by Chiappe (memorandum to RTI based on South Africa experience), "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."[11]

If test developers decide to use nonsense words in a language other than English, it is important to ensure that the syllabic structure makes sense. In English, nonsense words with a consonant-vowel-consonant (CVC) pattern, such as "wub" or "dod," are "legal" or consistent with the usual patterns of the language.

[11] *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."

V. EGRA Enumerator Training and Fieldwork

As noted in the introduction to Section IV, a week of training for enumerators who will be piloting the instrument is strongly recommended. This section is aimed at trainers who will be leading the training and overseeing the pilot fieldwork. It is expected that many of these trainers also will serve as enumerator supervisors in the field.

Ideally, all participants in the enumerator training should also have attended the adaptation and research workshop described in Section IV, although if necessary, the EGRA team may bring in additional enumerators or test applicators at this point to complement the Ministry staff and supervisor teams (see discussion of enumerator qualifications in Section III).

This section gives an overview of EGRA fieldwork and the initial piloting process, including lessons learned during the multiple pilots of the EGRA instruments. It is not a comprehensive supervisor manual, however; this toolkit assumes that such a manual would be developed by the technical assistance or country teams if it were determined that one was needed.

Discussion centers on these topics:

- Piloting the instrument
- Arriving at the school
- Selecting students and conducting the assessment
- Teaching lessons for the fieldwork

Piloting the Instrument

As described in Section IV, during the adaptation and research workshop, a Ministry team will have reviewed each of the instrument components and the underlying research.

At this follow-up training, participants will practice and pilot the instrument in several schools (roughly three to six, depending on the number of enumerators). Following training, a full-scale application will take place in the selected sample schools (see Annex B for sampling information). The instructions in this section regarding the school-based assessments apply to both the pilot test and the full-scale application. A sample agenda for this stage of the EGRA training is provided in Exhibit 10.

Exhibit 10. Sample Agenda: Enumerator Training and Pilot Fieldwork

Monday	Tuesday	Wednesday	Thursday	Friday
• Review underlying principles • Review draft instrument • Train in use of application, and practice	• Train in use of application, and practice • Review roles and responsi-bilities of supervisors and enumer-ators	• Pilot in 3 to 6 schools (determined by the number of teams being trained) • Enter data • Analyze results and modify instrument	• Print final version of the instrument • Train in use of application, and practice • Test for interrater reliability	• Prepare and pack materials • Finalize and review logistics

The objectives of the training include:

- Review underlying principles of EGRA in order to better understand the reasoning behind the instrument components.
- Solidify supervisor application and training practice and roles and responsibilities (development of the aforementioned supervisor manual may be useful in this case).
- Have the selected enumerators pilot the instrument in three to six schools (as a training exercise).
- Continue training enumerators in EGRA administration and scoring.
- Finalize the instrument and complete logistical preparations (printing, etc.).
- Notify sample schools of their selection, purpose of their assessment, and logistics needs (e.g., a separate, quiet room for administration of the instrument)

During the workshop, participants will need:

- copies of the complete draft enumerator and student instruments
- stopwatches or timers (if possible, find a kitchen timer that counts down from 1 minute)
- pencils with erasers, clipboards
- several laptops with Excel for data entry (one laptop per group of enumerators)

Workshop leaders should review and reinforce the skills tested in the instrument and the relationship of each component to instruction. The instrument should be carefully reviewed, paying close attention to clarity of the instructions for both students and enumerators. To the extent possible, student instructions should be consistent (including ways to encourage the child, number of examples, etc.) across each of the test components (refer to Exhibit 8 for a list of the components).

The number of enumerators to hire and train will depend on the number of schools to be visited and the timeframe for completion of the exercise. As the pilot in three to six schools is a training exercise, all enumerators selected for the overall administration should participate in the Week 2 workshop. At a minimum, enumerators and supervisors should visit schools in groups of four: one supervisor and three enumerators per team. In this way, supervisors can select students and circulate among enumerators during the testing period. Based on administrations of EGRA in several countries to date, it is estimated that results from 400 students are needed for each comparison group of interest (grade, school type, etc.; see Annex B for a detailed discussion on sample size). Thus, a simple national baseline comparing students by grade in grades 1 through 3 will require 1200 students. If additional comparison groups are required (e.g., rural versus urban, by grade), then 400 students are required for each group of comparison (in this example, 2400 students).[12] Exhibit 11 below summarizes the calculations for determining the number of schools and enumerators needed to conduct a survey reporting on results for grades 1 to 3. In several countries where EGRA has been conducted to date, Ministry staff members who participated in Week 1 of the workshop have been selected as supervisors for the duration of

[12] See Annex B for additional statistical support for this estimate using results from several EGRA applications.

the exercise. It is important to have more supervisors and enumerators than needed so as to be able to select from the best in both groups.

Exhibit 11. Estimates for Sampled Students, Schools, and Number of Enumerators

Comparison group of interest	Sample students	Sample schools (60 students per school)	Students per grade per school	Enumerators required (4 teams of 4)	Days of fieldwork (1 day per school per team)
Grades 1-3	1200	1200/60=20	20	16	5
Grades 1-3, Control vs. Treatment	2400	2400/60=40	20	16	10
Grades 1-3, Control vs. Treatment, Urban vs. Rural	4800	4800/60=80	20	16	20

Depending on the needs of the Ministry team and time available to complete the assessment, the number of enumerators can be increased or decreased. Thus, a sample of 40 schools can be completed in 10 days with 16 enumerators, or in 5 days with 32 enumerators. Application in these 40 schools should take place immediately following the pilot during Week 2.

Continuing with the example of the simple grade-comparison baseline of 1200 students, piloting should take place in at least four schools (one school per team of four), depending on the number of enumerators being trained. For the pilot, participants will need:

- copies of the final enumerator instrument
- one laminated set of student forms per enumerator (the same laminated forms will be used for each student that the enumerator tests)[13]
- stopwatches or timers (if possible, find a kitchen timer that counts down from 1 minute)
- pencils with erasers and clipboards
- pencils or other small school materials to give to students to keep in thanks for their participation

Testing for Interrater Reliability

Interrater reliability measures the degree to which different raters, or enumerators, agree in their scoring of the same observation. Interrater reliability is best used during the training process so as to improve the performance of the enumerators before they get to the field. It can also be used to aid in selection of the best-performing enumerators.

There are several ways to generate data for calculating interrater reliability, as follows.

[13] Because the student forms 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).

1. One enumerator assesses the student while another enumerator observes and scores at the same time. Enumerators then compare their scoring and discuss. Supervisors can also observe and score with each enumerator and discuss any discrepancies.

2. In a group setting, audio or video recordings of student assessment can be played while all enumerators score the assessment. Trainers can then collect the scoring sheets for review and comment (verification of coding and marking).

3. Adult trainers or enumerators can play the student role in small- or large-group settings and scoring sheets can be collected for review and comment. The benefit of this last scenario is that the adults can deliberately make several errors in any given subtest (e.g., skipping or repeating words or lines, varying voice volume, pausing for extended lengths of time to elicit prompts, etc.).

With all of these strategies, data should be collected for calculation of interrater reliability. Lead trainers should collect the scoring sheets for each subtask, input them into Excel, and calculate means and standard deviations. Those enumerators whose scoring results are greater than one standard deviation from the mean may require additional practice or support. If interrater reliability analysis reveals consistent poor performance on the part of an enumerator, and if performance does not improve following additional practice and support, that enumerator should not participate in the fieldwork.

Arriving at the School

Before departing for the schools, enumerators and supervisors should:

- Double-check all materials, including one copy of the laminated form of the student instrument per enumerator and sufficient copies of the enumerator instrument.
- Discuss test administration procedures and strategies for making students feel at ease, and role-play this exercise with one another.
- Verify that all administrators are comfortable using a stopwatch or their own watches.

Upon arrival at the school, the supervisor should introduce the team of enumerators to the school principal. In most countries, a signed letter from the Ministry will be required to conduct the exercise; the supervisor should present the letter (a copy of which should have been sent in advance, if possible; see an example of such a letter in Annex D), explain the purpose and objectives of the assessment, and thank the school principal for participating in the early grade reading assessment. The principal should be reminded that neither students nor teachers will be identified by name in the data collection process.

If planned, the school principal should be notified of the procedure for providing feedback to the school on the overall performance of the students. Finally, the supervisor should ask the principal if there is an available classroom, teacher room, or quiet place for each of the administrators to conduct the individual assessments. Enumerators should proceed to whatever space is indicated and set up two chairs or desks, one for the student and one for the enumerator.

Selecting Students and Conducting the Assessment

If recent and accurate data on student enrollment by school, grade and class are available at the central level prior to arrival at the school, a random number list can be used to generate the

student sample. As this is highly unlikely in nearly all low-income country contexts, the following procedure should be followed for selecting students in the school.

To assess 20 students per grade, the enumeration supervisor should conduct the following procedure for each grade (1, 2, and 3, one at a time).

1. Obtain the student register from each classroom or from the school principal, if available. It is a good practice to start with the selection of the first-grade children. Often, the later in the school day, the less able these small children are to focus and concentrate.

2. Count the total number of students registered in each class for that grade and add them up (e.g., Class A=40, Class B=30, Class C=50, Total=120).

3. Divide the total (120) by the number of students to be interviewed (in this example, 20 are to be selected from each grade [20 x 3] so the answer is 6).

4. Use this answer to count from the class lists and select each "X"th student to be part of the sample. In this example, the answer is 6, so students 6, 12, 18, 24, 30, and 36 on the list would participate from Class A; students 2, 8, 14, 20, and 26 from Class B; and students 2, 8, 14, 20, 26, 32, 38, 44, and 50 from Class C.

5. If a student is absent or refuses to participate, select the next number on the class list. If that student is absent or refuses, the following number should be selected. This will provide the sample of 20 students distributed across the grade 1 classes. Note the number of both refusals and absences in the school report.

6. Pull students out from their classes in small groups, 1 student per enumerator, so as to minimize the disruption to classes. Lead the students to the enumerators and introduce them by name. Note the size of the student's class and information on the student's age and/or birth date, and communicate this information to each of the enumerators at the start of the student interview.

7. Once the administrators have completed the assessment for all of the grade 1 students, repeat the same procedure as above for grades 2 and 3.

8. Ensure the administrators always have a student to assess so as not to lose time during the administration. To the extent possible, all interviews should be completed within the school day. If the school has only one shift and the assessment has not been completed before the end of the shift, find the remaining students and ask them to wait following the close of the school day. In this case, the school director or teachers should make provisions to notify parents that some children will be late coming home. This issue should be discussed in advance with enumerators and supervisors as to the most appropriate practice given local conditions.

Teaching Lessons for the Fieldwork

Throughout the training, participants should reflect on and share experiences from the piloting of the instrument. Instructions should be improved and clarified based on the experience of the enumerators in the schools. Actual fieldwork should take place immediately subsequent to the training. When possible, each team should have a car to transport materials and arrive at the sampled schools before the start of the school day. Experience to date has shown that

application of the EGRA requires about 15 to 20 minutes per child. This means that a team of three enumerators can complete about nine or 10 instruments per hour, or about 30 children in three uninterrupted hours.

Based on the work conducted by our local partner Centro de Investigación y Acción Educativa Social (CIASES) in Nicaragua, RTI has developed a supervisor manual to accompany the instrument and guide the data collection process. The manual can be found together with the most recent instruments on the EdData website (www.eddataglobal.org) under EGRA > Current EGRA Instruments.

VI. Analyzing the EGRA Data

This section covers some basic and low-technology approaches that the data entry and analysis team can use in working with the EGRA data.

Throughout the development of EGRA, RTI's approach has been to work with tools and approaches that are low cost and widely available. As statistical packages are quite expensive and require additional specialized training or self-teaching, EGRA counterparts have preferred to work with Excel. With that in mind, this section of the toolkit has been developed using Excel's "Data Analysis ToolPak" (an add-in available to most users of Excel) and Pivot Tables (see discussion below). When available and when they do not require much additional training in the use of the package itself, common statistical packages such as SPSS and Stata should be used because of their higher-order capacity for data analysis. This section addresses the following topics:

- Cleaning and entering data
- Using Excel to analyze data

A complete discussion of sampling weights is included in Annex B.

Cleaning and Entering Data

Experience with EGRA and other surveys suggests the following are important issues in ensuring good data entry (free of errors), and data cleaning.

1. Ensure that at least one key piece of identification data on each child is included on every sheet of the questionnaire or assessment form, in case questionnaires or assessment forms become separated in transport.

2. Ensure that all data forms are checked for completeness and consistency at the end of each day, ideally by someone other than the person carrying out the assessment. This implies a reasonable ratio of supervisors to assessors, so that the supervisors can get through the forms at the end of the day. Alternatively, assessors can check each other's work at the end of the day.

3. Given the relatively small sample sizes used in EGRA assessments, data entry has been done in Excel in many cases. This maximizes transparency and ease in sharing the data. Excel can be used to create simple functions and comparisons that allow automatic consistency and range checks on the data, to detect and prevent data-entry errors. Other methods of data entry are possible, naturally, but for the sample sizes being considered for most EGRA assessments, Excel is sufficient. A Microsoft Access-based data entry interface system has been developed and is being tested in several countries. This standardized data entry system greatly reduces data entry error and can be programmed to generate simple reports.

4. However, as noted above, it is likely that once data are entered with Excel, one may want to transfer the data to a statistical package, such as SPSS or Stata, for actual analysis. Thus, it is important to enter the data with a great deal of care so as to protect record integrity. That is, it is important to make sure that the data for a given child are

carefully entered into a single row, and to take great care not to mix up rows. Mixed or multiple rows are particularly dangerous if the data are sorted electronically. Excel has very weak record integrity capacity, so data entry and analysis personnel must take a great deal of care when manipulating data.

5. As noted in Section IV, in coding the data, it is extremely important that data entry personnel record the answers correctly, and have a strategy for coding that differentiates among the following types of responses: (a) question not asked, (b) student did not know, (c) student would not answer or could not answer, and (d) a true zero (completely incorrect answer).

6. Similarly, if the data are to be shared with others, or are to be imported into a statistical package, it is important to create variable names that are complete and mnemonically useful (e.g. CLPM for Correct Letters per Minute).

7. For some key variables, such as fluency (correct words per minute), the data as they come directly from the assessment forms will denote the time a given task took, and the number of words or letters read. That is, even though the task is timed at 1 minute, a few children may finish the task in less than 1 minute, and in some cases more than 1 minute may be allowed. It is therefore recommended that both values, namely the number of words or letters read *and* the time in seconds, be entered into the database, but that the Excel functionality be used to create the "per minute" variable. Thus, one would enter that a child read 55 words correctly in 40 seconds, and then use an Excel formula to calculate that this means the child read at 82.5 words per minute, where the formula would be correct words) seconds x 60 (see discussion above).

8. A codebook or data dictionary should be created by the data analysis team to describe each variable name. This could most easily be located in a separate page in the Excel worksheet, and would contain information similar to that in Exhibit 12:

Exhibit 12. Sample Codebook Entries

Variable name	Variable description	Coding notes
Clread	Correct letters read	Blank means that the task was not continued
Clseconds	Time in seconds to correctly read letters	Blank means that the task was not continued
Clpm	Fluency in correct letters read per minute (created variable)	Zero means that no words were read as the task was not continued
Cwctread	Correct words in connected text read	Blank means that the task was not continued
Cwctseconds	Time in seconds to read connected text	Blank means that the task was not continued
Cwpmct	Correct words per minute in connected text (created variable)	Zero means that no words were read as the task was not continued

9. In entering data and creating variables, the data analysis team should create two variables for each of the important concepts, such as correct words per minute in connected text, or correct letters per minute. For calculating averages, for example, it may be useful to create one average that includes only the children who attempted the task and read enough to be allowed to continue the task, and one that also includes those who simply did not read enough to be recorded. If only the former children are included, it creates a somewhat distorted picture of the school, since it exaggerates how well children are reading. But if children who cannot read at all are included in the average, it does not give a good sense for the reading fluency of those who can indeed read.

   There is no simple solution to this problem, particularly since the line between "can read" and "cannot read" is actually somewhat arbitrary, as can be seen in the fact that the calculated fluency will typically range very widely and will show many cases close to zero. A good solution, therefore, is to enable both sorts of calculations: one average that includes the children who cannot read at all as having a fluency of 0, and one average that excludes the children who were judged nonreaders, and thus includes only those with fluency greater than 0. If a program such as Stata or SPSS is used, the matter is simple, as there are simple commands for excluding cases with a value of 0 in the variable whose average is being calculated.

Data Analysis: Using Excel to Analyze Data

Most or all of the analyses needed to produce a basic report on EGRA results can be done with Excel. This section suggests how most of the calculations can be done. In addition, sampling weights should be used for proper weighting of the results (see Annex B for additional information).

The following suggestions should enable all of the basic analyses needed.

To protect data integrity, it is strongly suggested that as much of the analysis as possible be done using the Excel "Pivot Table" facility. This allows the data analysis team to calculate averages, for the key variables, according to any needed subsets of the data, such as by age, by grade, by gender, or by school. For any variable, and for any subgroup, one can easily calculate the usual results such as the mean, the count or number of cases, and the standard deviation.

Results typically of interest will include the average correct words (or letters) per minute, broken down by age or grade or some other factor. A typical example, generated using Excel, would be as follows:

	Grade		
Variable	1	2	3
Correct letters per minute	12.8	25.4	36.1
Correct letters per minute, excluding nonreaders	19.8	28.6	37.7
Correct familiar words per minute	1.2	2.3	4.3
Correct familiar words per minute, excluding nonreaders	6.2	8.3	9.2
Correct words per minute, connected text	2.2	4.0	9.2
Correct words per minute, connected text, excluding nonreaders	11.0	11.6	17.3

It may be of interest to see whether there are differences in the results by grade or age or gender. For example, in this case, there seems to be some grade progression in the fluency of letter naming: a gain of about 10 letters per grade. But how sure can we be that differences by grade, by age, or by gender are significant?

A reasonably rigorous sense of how significant the differences are by grade (or age, gender, public-private, urban-rural, or any other attribute) can be derived by asking Excel, via the Pivot Table command, for the means, standard deviations, and counts, by grade, and then computing a simple confidence interval. A simple example, for the correct letters per minute variable above, is as follows:

	Grade		
	1	2	3
Count	419	389	392
Average correct letters per minute	12.8	25.4	36.1
Standard deviation	18.2	21.9	23.2
Standard error	0.89	1.11	1.17

A table generated in this way, however, does not yet give the confidence intervals. The table tells us that there is substantial variability *within* grades; some children perform much better than others. Since most children are, of course, not at the average, the "average child" is at some distance, up or down, from the average of the children. The standard deviation tells us that, loosely speaking, the average child's performance can range as much as 20 or so words above or below the average of the children's performance. In other words, the standard deviation is a measure of the average difference of each individual from the average.

This variability is interesting for its own sake, since it tells us something about inequality or unevenness of performance. But it can also be used to give a sense of how statistically reliable the averages are. The intuition as to why the standard deviation might guide us as to how reliable the averages are—or how reliable the differences between the averages are—is based on the notion that if there is a lot of variation within the grades, then maybe the differences observed between the grades are accidental. Since all we have is a sample, not the actual population, we may be mistaken as to the correct letters per minute that children in the population can read. Any given sample could have erred on the high side or on the low side, and the greater the variability, the greater the likelihood that a sample could err on either the high side or the low side. But how likely is it that one could have erred so much as to come to the wrong conclusion that there is some grade progression when there is actually none?

It is possible to answer this question with some rigor by building something called "confidence intervals." To do this takes two steps, using very simple Excel commands or formulae. First, compute the "standard error," which is the standard deviation divided by the square root of the count. Thus, for example, the standard error for correct letters per minute in grade 1 is 0.89 or 18.2)√419. Second, and using a simplified rule of thumb, add twice the standard error to the mean to get an "upper bound" on the confidence interval, and subtract twice the standard error from the mean to get a "lower bound" on the confidence interval. This can be tabulated, again using Excel, as follows.

	Grade		
	1	2	3
Average correct letters per minute	12.8	25.4	36.1
Lower bound of confidence interval	11.0	23.1	33.7
Upper bound of confidence interval	14.6	27.6	38.4

In this table, for example, 14.6 is just 12.8 + 2 * 0.89, or the average plus twice the standard error. This table should be interpreted as follows: "The average correct letters per minute in grade 3 is 36.1 in our *sample*, and we can be 95 percent confident that the underlying average in the *population* is somewhere between 33.7 and 38.4." If the upper bound of one grade, say grade 2, is lower than the lower bound of the next grade, say 33.7, then we can be quite confident that there is a real grade progression.[14]

Sample Size

At this point a discussion of sample size is possible. It may seem that sample size should have been discussed in more detail previously. However, it is easier to understand references to sample size after one has seen the sorts of confidence intervals that can be generated using real examples of sample sizes. In the table above, we have seen that confidence intervals for, for example, correct letters per minute, tend to be plus or minus 1.8 around the average, if the sample size is some 400 children per grade; and that this is enough to detect progression between grades. This is dependent, of course, on the standard deviation. Based on EGRA experiences, we are beginning to discover that the standard deviation is reasonably constant across countries.

The basic results for other concepts such as correct words per minute in connected text—based on analysis of data in Peru, Jamaica, Kenya, The Gambia, and others—do not differ that much from each other, showing standard deviations around 15 to 20 on the high side. The following table shows the results for connected text from Kenya.

	Kiswahili	English
Count	400	400
Mean correct words per minute, connected text	8.7	9.3
Standard deviation	13.1	14.2
Lower bound of confidence interval	7.4	7.9
Upper bound of confidence interval	10.0	10.7

Based on these kinds of results, sample sizes of approximately 400 children per group (gender, grade, language, public-private, urban-rural) whose values are worth considering disaggregating seem adequate. Naturally, some 400 children are needed for any *combination* (male-urban would require some 400, as would female-urban, male-rural, and female-rural; thus, distinguishing by gender and locality would require a sample size of some 1600, whereas a simple baseline per grade would require only 400 per grade). For a detailed discussion on sampling size and weighting procedures, please see Annex B.

[14] See Annex B for a discussion of the balancing act required in decisions about sampling approach, sample size, and desired level of precision.

VII. Using EGRA: Implications for Policy Dialogue

In classrooms around the world, there is a disconnect between instructional practices and student performance. Teachers frequently write long paragraphs on the chalkboard for students to copy; in these same classrooms many students cannot recognize all of the letters of the alphabet. Many of the advances in reading instruction demonstrated to be effective in numerous research studies are not being used by teachers. In fact, in many countries, very little systematic teaching of reading is going on. In many cases, the five components of effective reading instruction outlined by the National Reading Panel (2000) (phonemic awareness, phonics, fluency, vocabulary, and comprehension) are nowhere to be found. Although explanations for this disconnect are myriad, at least part of the problem lies in the gap between the research community and what is being taught in teacher training programs in low-income countries.[15]

As discussed in Section II, one of the primary purposes of EGRA is to diagnose, at the system level, areas for improvement in early reading instruction. The following section provides a brief overview of policy and instruction issues to be addressed by ministries and donors in using EGRA results for this purpose.

Using Results to Inform Policy Dialogue

The ultimate purpose of EGRA is to inform changes in instruction. In our experience, thus far, the impact on policy dialogue to inform instruction seems to have two separate steps.

Influencing Policy Makers and Officials

First, the results tend to concern policy makers and officials. One of the virtues of EGRA is that the science behind it seems to correspond fairly well to the average citizen'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 are what most citizens intuitively think of as reading. Thus, being able to report that children cannot recognize letters, or can read them only extremely slowly, is something that most citizens can understand. The utilization of audio or video recordings that dramatize the differences between a very poor reader (a child reading at, say 10–15 words per minute, with no comprehension) and a better reader (a child reading at, say, 60 words per minute, with comprehension) is instantly obvious and dramatic. (For an example of such a video, developed by DFID and the World Bank in Peru, see www.eddataglobal.org, main page).

Furthermore, citizens and officials, particularly those who apply the EGRA test themselves (or simply ask children to read to them), quickly develop a sense that children are not reading, and communicate this to other officials. Officials in various countries 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 the kinds of concerns officials already have been expressing.

In some contexts, reactions to an EGRA-type reading assessment are not as straightforward. Some commentators, in some countries, question the usefulness of oral reading fluency as a

[15] Under the EQUIP1 program financed by USAID, the International Reading Association and the American Institute for Research are currently reviewing a sample of national curricula and instructional materials to assess the emphasis (or lack thereof) on reading instruction.

marker or precursor indicator of general learning, or even of reading. This is why it is important to have access to the background literature that explains the issues, some of which is referenced in this toolkit. Other useful references can be found at www.reading.org, www.nationalreadingpanel.org, and http://dibels.uoregon.edu/.

> "To prevent reading difficulties, children should be provided with:
> - Opportunities to explore the various uses and functions of written language and to develop appreciation and command of them.
> - Opportunities to grasp and master the use of the alphabetic principle for reading and writing.
> - Opportunities to develop and enhance language and meta-cognitive skills to meet the demands of understanding printed texts.
> - Opportunities to experience contexts that promote enthusiasm and success in learning to read and write, as well as learning *by* reading and writing.
> - Opportunities for children *likely to experience* difficulties in becoming fluent readers to be identified and to participate in effective prevention programs.
> - Opportunities for children *experiencing* difficulties in becoming fluent readers to be identified and to participate in effective intervention and remediation programs, well integrated with ongoing good classroom instruction."
>
> Snow, C.E. et al. (1998), *Preventing reading difficulties in young children* (p. 278)

In other cases, a perception seems to exist 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, but that, indeed, the effort is not based on the assumption that reading is all that matters.

In general, any attempt to measure 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, tend to see the virtue in EGRA; whereas some curricular or reading theoreticians seem to 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, and 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.

In focusing the attention of policy makers and officials on the subject, it is useful to be able to benchmark the results in some way. Two ways of benchmarking have been found useful in EGRA exercises in any given country: a comparison to international standards or goals of some sort; and some analysis of the performance of schools in the country in question. Exhibit 13 below shows actual average results in a recent EGRA exercise, one possible international benchmark, and one possible national benchmark.

Data in the last column, namely the international benchmarks, were taken from the Dynamic Indicators of Basic Early Literacy Skills.[16] Additional indicators for connected text for U.S. students can be found in Annex A.

[16] The DIBELS can be found at http://dibels.uoregon.edu/benchmark.php.

Exhibit 13. Example of Possible Benchmarking Exercise

	Average across children in a given country, middle of year, grade 2	Maximum school-level average in the same country, grade 2	Developed-country benchmarks for comparison purposes
Correct letters per minute	22.7	41	40 by end of kindergarten year*
Correct nonsense words per minute	7.5	25	50 in middle of grade 1
Correct words per minute	11.4	36	20 in middle of grade 1
Comprehension score	0.4	2	NA

* This skill is not even tracked or benchmarked past kindergarten, on the assumption that it is mastered in kindergarten. Note the low level of the average, for grade 2, for the case in the table.

The most important item to benchmark—because its predictive power over other skills is highest—is connected-text fluency (correct connected words per minute). For this, most scholars converge on the idea that by the end of grade 2, children learning to read in English ought to be reading at about 60 correct words per minute, respectively. Based on our experience in approximately 10 countries to date, for a poor country with linguistic complexity or particularly difficult orthographies, these benchmarks could perhaps reasonably be relaxed to something like 45 correct words per minute. Based on these results, researchers could then specify the proportions of children who achieve foundation-level accuracy and fluency of reading in addition to the average grade at which children "break through" to reading literacy. The "breakthrough" grade could then be the grade at which 90 percent of children are meeting some standard or benchmark.[17]

Most importantly, a country can set its own benchmarks by looking at performance in schools that are known to perform well, or 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 text comprehension questions also indicate whether children are reading "with understanding." Again, a criterion needs to be specified based on empirical results and examination of the distribution (e.g., 75 percent of questions answered correctly) for each country. The proportion of children who achieve foundation reading literacy with fluency and comprehension can then be specified. The assessment can provide analytic and diagnostic information about reading progress in particular schools/classrooms (with the right sampling strategy) and can be reported back to teachers at the classroom level. For example, progress may be limited by poor comprehension, lack of fluency, lack of letter-sound knowledge, inability to decode, narrow scope of sight vocabulary, or combinations of these features. This information can be returned to teachers, school directors, and school supervisors as an

[17] If, particularly in the first few cases where these instruments are tried, the grades at which the instrument is tested do not cover a grade in which, for example, 90 percent of the children are breaking through to reading literacy, regression and extrapolation techniques can be used to estimate this grade. Subsequent implementation then will have to be sure to include this grade.

indication of where the balance of instruction might be shifted so as to improve learning outcomes.

Changing Reading Instruction

Second, at least when things go right, concern seems to turn to a realization of the need to change instruction in reading in the early grades. So far, two countries where EGRA has been tried, The Gambia and South Africa, spontaneously developed fairly extensive materials for training teachers in improved teaching of reading. In Kenya, donor activity and collaboration with the Aga Khan Foundation led to the development of a set of lesson plans to improve the teaching of reading in the early grades. Annex E includes the outline of a week of collaboration among a reading expert, the Aga Khan Foundation, and local government officials to develop lesson plans.

Sample lesson plans and strategies for teaching the foundation aspects of early reading—namely phonemic awareness, phonics, fluency, vocabulary and comprehension—are available from a number of sources. Examples include teacher handbooks such as those of Linan-Thompson and Vaughn for teaching native English (2004) and English language learners (2007). Such materials can be used to inform locally developed lesson plans (see an example in Annex F). Countries can also develop their own suggested lesson plans using the EGRA results to identify those areas that need improvement. Examples of the lesson plans from Kenya, The Gambia, and South Africa can be found at www.eddataglobal.org, Documents and Data.)

Using Data to Report Results to Schools

To date, applications of EGRA have been used primarily to generate discussion at the national level and to spur Ministries into action. Complementing this promotion of awareness is the reporting of results to schools and teachers. To reiterate a statement from an earlier section: In no case should individual students or teachers be identified in reporting to schools, as the measure is not meant be used as a high-stakes accountability tool. That said, some form of give-back to schools as a means of thanking them for their participation is usually welcome.

To report results to schools, analysts should create a simple one-page summary of results, including reporting by grade and gender, for each individual school. Average results for schools with similar characteristics and means for the entire sample can be shared as well. This reporting should be accompanied by explanations as to how each subtest relates to instruction and what teachers can do to improve student results. Sample lesson plans and suggested activities should also be shared with schools.

References

Abadzi, H. (2006). *Efficient learning for the poor*. Washington, DC: The World Bank.

Abadzi, H., Crouch, L., Echegaray, M., Pasco, C., & Sampe, J. (2005). Monitoring basic skills acquisition through rapid learning assessments: A case study from Peru. *Prospects, 35*(2), 137-156.

AskOxford.com. (n.d.). *Ask the experts: Frequently asked questions, Words*. Retrieved April 2008, from http://www.askoxford.com/asktheexperts/faq/aboutwords/frequency?view=uk

Carver, R. P. (1998). Predicting reading level in grades 1 to 6 from listening level and decoding level: Testing theory relevant to the simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 10*, 121-154.

Catts, H. W., Hogan, T. P., & Fey, M. E. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *Journal of Learning Disabilities, 36*, 151-164.

Center for Global Development. (2006). *When will we ever learn? Improving lives through impact evaluation*. Retrieved January 2007, from www.cgdev.org/files/7973_file_WillWeEverLearn.pdf

Chabbott, C. (2006). Accelerating early grades reading in high priority EFA Countries: A desk review. From http://www.equip123.net/docs/E1-EGRinEFACountriesDeskStudy.pdf

Chiappe, P. (2006). *Rapid assessment of beginning reading proficiency: A proposed research design*. Unpublished manuscript.

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.

Children's Literacy Initiative. (2000). *Dictation task*. Retrieved March 2007, from http://www.cliontheweb.org/pd_asmntsamp2.html

Clay, M. M. (1993). *An observation survey of early literacy achievement*. Ortonville, MI.: Cornucopia Books.

Crouch, L. (2006). *La fluidez lectora como indicador operacional [Reading fluency as an operational indicator]*. Unpublished manuscript, Washington, DC.

Crouch, L., & Winkler, D. (2007). *Governance, management and financing of Education for All: Basic frameworks and case studies*. Unpublished manuscript.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can

depend on how comprehension is measured. *Scientific Studies of Reading, 10*(3), 277-299.

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.

Department for Children, Schools and Families. (2006). *The new conceptual framework for teaching reading: The "simple view of reading." Overview for literacy leaders and managers in schools and early years settings*. Retrieved August 2007, from www.standards.dfes.gov.uk/eyfs/resources/downloads/paper_on_searchlights_model.pdf

Dolch, E. (1948). *Problems in reading*. Champaign, IL: The Garrard Press.

Educational Testing Service. (2005). *Dictation assessment*. Retrieved March 2007, from http://www.pathwisestore.com/index.asp?PageAction=VIEWPROD&ProdID=161

Ehri, L. C. (1998). Word reading by sight and by analogy in beginning readers. In C. Hulme & R. M. Joshi (Eds.), *Reading and spelling: Development and disorders* (pp. 87-111). Mahwah, NJ: Erlbaum.

Espin, C., & Foegen, A. (1996). Validity of three general outcome measures for predicting secondary students: Performance on content-area tasks. *Exceptional Children, 62*, 497-514.

Espin, C., & Tindal, G. (1998). Curriculum-based measurement for secondary students. In M. R. Shin (Ed.), *Advances applications of curriculum-based measurement*. New York: Guilford Press.

Filmer, D., Hasan, A., & Pritchett, L. (2006). *A millennium learning goal: Measuring real progress in education*. Washington, D.C.: The World Bank.

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.

Good, R. H., III, Simmons, D. C., & Smith, S. B. (1998). Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills. *School Psychology Review, 27*(1), 45-56.

Gough, P. B. (1996). How children learn to read and why they fail. *Annals of Dyslexia, 46*, 3-20.

Gough, P. B., & Tunmer, W. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*, 6-10.

Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *International Reading Association*, 636–644.

Hirsch, E. D., Jr. (2003). Reading comprehension requires knowledge--of words and the world. *American Educator* (Spring), 1-44.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*, 127-160.

International Literacy Institute, & UNESCO. (2002). *Towards guidelines for the improvement of literacy assessment in developing countries: Conceptual dimensions based on the LAP project*. Unpublished manuscript, Philadelphia, PA.

International Reading Association. (2007). *Teaching reading well: A synthesis of the International Reading Association's research on teacher preparation for reading instruction*. Retrieved January 2008, from www.reading.org/resources/issues/status.html

Joshi, R. M., & Aaron, P. G. (2000). The component model of reading: Simple view of reading made a little more complex. *Reading Psychology, 21*, 85-97.

Kame'enui, E. J., Fuchs, L., Francis, D. J., Good, R. H., III, O'Connor, R. E., Simmons, D. C., et al. (2006). The adequacy of tools for assessing reading competence: A framework and review. *Educational Researcher, 35*(4), 3-11.

Kaminski, R. A., Good, R. H., III, Baker, D., Cummings, K., Dufour-Martel, C., Fleming, K., et al. (2006). *Position paper on use of DIBELS for system-wide accountability decisions*. Retrieved January 2007, from www.cde.state.co.us/action/CBLA/Accountability_2006-11-16.pdf

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). Department of Psychology, Florida State University.

Moats, L. (2000). *Speech to print: Language essentials for teachers.* Baltimore, MD: Paul H. Brookes.

Moats, L. (2004). *Language essentials for teachers of reading and spelling*. Frederick, CO: Sopris West Educational Services.

Mullins, I., Martin, M., Gonzalez, E., & Chrostowski, S. (2004). *TIMSS 2003 international mathematics report.* Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education.

National Institute of Child Health and Human Development. (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)*. Retrieved August 2007, from http://www.nichd.nih.gov/publications/nrp/upload/report_pdf.pdf

National Literacy Panel. (2004). *National Literacy Panel on Language Minority Children and Youth: Progress report*. Retrieved August 2007, from http://www.cal.org/natl-lit-panel/reports/progress.pdf

Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003*. Retrieved January 2008, from http://www.oecd.org/document/55/0,3343,en_32252351_32236173_33917303_1_1_1_1,00.html

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.

Peereman, R., & Content, A. (1999). LEXOP: A lexical database providing orthography-phonology statistics for French monosyllabic words. *Behavioral Methods, Instruments and Computers* (31), 376-379.

Pratham. (2005). *Annual Status of Education Report (ASER): Final report*. Retrieved April 2006, from http://www.pratham.org/aserrep.php

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. (1999). Phonological recoding and orthographic learning: A direct test of the self-teaching hypothesis. *Journal of Experimental Child Psychology, 72*, 95-129.

Share, D. L., Jorm, A., Maclearn, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Education Psychology, 76*, 1309-1324.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: Committee on Preventing of Reading Difficulties in Young Children and National Academy Press.

Sprenger-Charolles, L., Colé, P., & Serniclaes, W. (2006). *Reading acquisition and developmental dyslexia.* New York, NY: Psychology Press.

Sprenger-Charolles, L., Siegel, L., Béchennec, D., & Serniclaes, W. (2003). Development of phonological and orthographic processing in reading aloud, in silent reading and in spelling: A four year longitudinal study. *Journal of Experimental Child Psychology, 84*, 194-217.

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.

Torgesen, J. K. (1998). Catch them before they fall: Identification and assessment to prevent reading failure in young children. *America Educator/American Federation of Teachers, 22*, 32-39.

Wagner, D. A. (2003). Smaller, quicker, cheaper: Alternative strategies for literacy assessment in the UN Literacy Decade. *International Journal of Educational Research, 39*.

World Bank. (2007). *Por una educación de calidad para el Perú: Estándares, rendición de cuentas, y fortalecimiento de capacidades [Quality education for Peru: Standards, accountability and strengthening skills]*. Washington, DC: The World Bank.

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.

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.

Zeno, S. M., Ivenz, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

Zieger, J., & Goswami, U. (2005). Reading acquisition, developmental dyslexia and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 13*(1), 3-29.

Annex A. English Oral Reading Fluency Norms for the United States

TABLE 1
Oral reading fluency norms, grades 1-8

Grade	Percentile	Fall WCPM	Winter WCPM	Spring WCPM
1	90		81	111
	75		47	82
	50		23	53
	25		12	28
	10		6	15
	SD		32	39
	Count		16,950	19,434
2	90	106	125	142
	75	79	100	117
	50	51	72	89
	25	25	42	61
	10	11	18	31
	SD	37	41	42
	Count	15,896	18,229	20,128
3	90	128	146	162
	75	99	120	137
	50	71	92	107
	25	44	62	78
	10	21	36	48
	SD	40	43	44
	Count	16,988	17,383	18,372
4	90	145	166	180
	75	119	139	152
	50	94	112	123
	25	68	87	98
	10	45	61	72
	SD	40	41	43
	Count	16,523	14,572	16,269
5	90	166	182	194
	75	139	156	168
	50	110	127	139
	25	85	99	109
	10	61	74	83
	SD	45	44	45
	Count	16,212	13,331	15,292
6	90	177	195	204
	75	153	167	177
	50	127	140	150
	25	98	111	122
	10	68	82	93
	SD	42	45	44
	Count	10,520	9,218	11,290
7	90	180	192	202
	75	156	165	177
	50	128	136	150
	25	102	109	123
	10	79	88	98
	SD	40	43	41
	Count	6,482	4,058	5,998
8	90	185	199	199
	75	161	173	177
	50	133	146	151
	25	106	115	124
	10	77	84	97
	SD	43	45	41
	Count	5,546	3,496	5,335

WCPM: Words correct per minute
SD: Standard deviation
Count: Number of student scores

Source: Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *International Reading Association,* 636–644. Texts are designed to be appropriate for each grade level.

Annex B. Sample Size Considerations in Early Grade Reading Assessment

Introduction

This note sets out basic sample size considerations applicable to Early Grade Reading Assessment (EGRA) samples. It is designed to inform Ministry staff, donors, or other actors interested in setting up an EGRA on sampling size requirements and calculations. The note assumes familiarity with basic statistics, highlights only the issues that are not normally found in a standard undergraduate statistics textbook, and does not define common statistical terminology. As discussed in Section VI, it is possible to conduct the majority of these calculations using Excel; however, Stata or SPSS is preferred (for data preservation and because the calculations can be programmed using syntax files). This annex refers to calculations using both Stata and Excel.

Broadly speaking, sample sizes are determined by a set of factors that include the characteristics of the data and a series of researcher-selected assumptions. In the case of EGRA, the sample size is determined by the variability of learners' performance on past EGRA assessments, the level of precision that the researcher would like to see in the resulting data, and the sampling approach applied.

The greater the variability of the learners' performance, the greater the required sample size. If, for example, one goes to a school where all students are reading at exactly the same fluency level, one would only need to sample one student to calculate an accurate estimate of the average reading fluency in that school. Unfortunately, actual variability cannot be known in advance, when one is planning the sample size. We know that there is always *some* variability among students. One way to develop sample size estimates for a new case (new country, new region) is to look at other cases.

As noted, in addition to the data's characteristics, the precision that the researcher would like to see in the results also will have an impact on the sample size. To understand the basic issue here, one has to realize that a sample can only give sample-based estimates of the underlying value for the total population. For example, a sample's average reading fluency is only an estimate of the reading fluency of the underlying population. After all, a sample is just a sample, and any other sample could give a different value. We then want to know how precisely any given sample is estimating the value in the underlying population. How precisely does our sample-based estimate of, say, reading fluency, estimate the reading fluency of the underlying population? This is a key issue. The advantage of sampling is that it can save cost, relative to evaluating the entire underlying population, but if the estimates are too imprecise, this advantage is not worth much. Greater precision generally requires a larger sample size, but how much larger?

To begin to get at this issue, it is important to realize that the notion of "precision" has two aspects to it. First, what is the size of the range in which the learners' performance score could fall? In statistics, it is traditional to say something like "our sample estimates that children read 50 words per minute, and therefore the children in the underlying population are most likely reading 50 words per minute plus or minus 5 words per minute." A less precise estimate would say "our sample estimates that children read 50 words per minute, and therefore the children in the underlying population are most likely reading 50 words per minute plus or minus 20 words

per minute." The notion of "plus or minus" is called a confidence interval (CI), and the actual value of the "plus or minus" is called the *width* of the confidence interval. The smaller the width, the more precise the results. The second issue is already hinted at above: One can say, "Therefore the children in the underlying population are 'most likely' reading 50 words per minute plus or minus 10 words per minute." But how likely is "most likely?" That is the second aspect of precision: how confident do we want to be that we have captured the real value of the learner's performance: 90% confident, 95% confident, or 99% confident? In statistical terms this is known as the *confidence level.* An intuitive way to interpret this, in the context of sampling, is to realize that to say "we are 99% confident that the population average is 50 words per minute plus or minus 5" is more or less equivalent to saying "there is only about a 1% chance that any given sample would have given us a sample average of 50 if the underlying population mean were outside the range of 45 to 55." So, we have a *confidence level*, and a *Width* of the confidence interval. We can state with a given level of precision how well our sample average estimates the population average.

The approach used in sampling or selecting students for participation in the EGRA will also have an impact on the sample design. We discuss this in more detail below.

Given that larger sample sizes can accommodate large levels of variability in student performance and provide more precise results, one might conclude that a large sample should always be selected. Unfortunately, large samples are very expensive. The researcher must select a sample that is large enough to provide reliable data while not requiring excessive funding levels. Also, any school-based assessment exercise interrupts school procedures, takes time away from lessons, and is an imposition. There are, therefore, certain ethical considerations in keeping the sample size relatively small.

As discussed in the initial sections of this toolkit, the purpose for which EGRA has been designed is to *conduct a system-level diagnosis for a given level.* Based on this purpose, the discussion below centers on strategies for drawing a nationally representative sample. Throughout this annex, the discussion is framed in terms of the minimum sample size needed for reporting results for one grade. That said, most of the EGRA exercises have been developed for assessing in multiple early grades (for example, grades 1–3 or 2–4). The calculations below, therefore, apply for the minimum sample size needed for each grade of interest: If countries are interested in testing skills in three grades, the sample should be drawn, using the parameters below, for each of those three grades.

Fortunately, sufficient experience has accumulated with EGRA to enable a reasonable determination of recommended sample sizes. Experiences in Peru, Pakistan, The Gambia, Senegal, and Jamaica, as well as two experiences in Kenya, can now be used to underpin sample size discussion and recommendations. Table 1 contains key background information helpful in determining a sample size, for all the countries named above. It uses reading fluency in connected text (the paragraph reading segment) as the key variable of interest. Country names are not provided, as these data are given to illustrate within-country patterns of differences between grades and genders (to pick two attributes) and variability in general. These data are not meant to be used for cross-country comparisons of reading fluency.

Table 1. Key reading fluency estimates across countries and grades in various EGRA efforts

Country	Grade 1	Grade 2	Grade 3	Grade 4	Total	Gender difference for all grades	Judgmental country weight
Country 1							6
Male	2.4	17.8	28.7		16.2		
Female	3.3	17.0	35.6		18.6		
Grade-wise average across genders	2.9	17.4	32.4		17.5	-2.4	
Grade-wise standard deviation	5.9	17.4	23.5		21.0		
Average inter-grade gain		14.8					
Country 2							10
Male	1.9	4.3	9.9		5.3		
Female	2.4	3.6	8.6		4.8		
Grade-wise average across genders	2.2	4.0	9.2		5.1	0.5	
Grade-wise standard deviation	9.3	12.4	19.9		14.8		
Average inter-grade gain		3.5					
Country 3							8
Male		59.8	66.8		63.5		
Female		58.3	78.7		68.3		
Grade-wise average across genders		59.0	73.1		66.1	-4.9	
Grade-wise standard deviation		46.8	48.1		47.9		
Average inter-grade gain		14.1					
Country 4							10
Male	23.2	30.4	50.3	68.3	45.8		
Female	28.2	36.2	58.1	90.2	56.1		
Grade-wise average across genders	25.3	33.1	53.9	78.1	50.5	-10.3	
Grade-wise standard deviation	30.5	34.4	39.2	46.5	43.2		
Average inter-grade gain		17.6					
Country 5							6
Grade-wise average across genders	9.2	29.3					
Grade-wise standard deviation	16.9	30.7			27.4		
Average inter-grade gain		20.1					
Country 6							3
Male	6.8	30.0	97.2	100.5	59.6		
Female	7.3	31.5	44.4	68.5	37.5		
Grade-wise average across genders	7.0	30.8	69.3	85.0	48.2	22.1	
Grade-wise standard deviation	15.0	39.0	79.3	68.9	64.0		
Average inter-grade gain		26.0					
Country 6 – special case							10
Male		11.5					
Female		11.2				0.4	
Grade-wise average across genders		11.4					

Country	Grade 1	2	3	4	Total	Gender difference for all grades	Judgmental country weight
Grade-wise standard deviation		16.2					
Average inter-grade gain			NA				
Across-country averages							
Average fluency by grade across all countries	10.5	25.0	43.7	79.7			
Average standard deviation of fluency by grade across all countries	16.5	26.5	36.6	51.6			
Average standard deviation across all grades (using all data points above, not averages across countries as in the row immediately above)			29.2				
Average inter-grade gain across all countries			14.1				
Average gender difference across all countries			-3.4				

NA = Not applicable.

Notes: All averages weighted. The approximate country weight is based on the authors' judgment of total importance given the sample size and the rigor with which the sample was selected.

Sources: Calculated from various countries' EGRA databases.

In what follows, we explain how sample sizes can be calculated given existing EGRA data, assumed confidence intervals, confidence levels, and a set sampling approach.

To recall the definition from above, a *confidence interval* is a range of values (as opposed to one number) used to estimate a population parameter (such as the average reading fluency of the underlying population) based on a sample estimate (the sample-based estimate of fluency). The narrower or the smaller the width of the confidence interval, the more reliable or precise the results will be. The size of the confidence interval that the researcher sets will depend on the characteristics of the variable being studied. A common approach to suggesting an appropriate width for a confidence interval is to look at ranges of variation across key attributes of the learners, such as grade or gender, and to suggest that the confidence intervals be narrow enough to allow for a distinction in performance along those key attributes. It is reasonable to ask, for example, that the confidence intervals for different grades not overlap each other. From the last panel of Table 1 above, it is clear that the average difference between grades across all countries is 14.1, or 14 to take a rounded number. This, then, seems a reasonable width upon which to base estimates of sample size.

Confidence intervals are associated with specific *confidence levels*. The confidence level tells us the probability that the confidence interval contains the true population parameter (the mean reading fluency). The greater the confidence level, the greater the level of precision. Researchers normally assume confidence levels of 90%, 95%, and 99%, with 90% considered somewhat marginal.

Sampling Approach

As mentioned previously, the applied sampling approach will also impact the sample size requirements. Other things being equal, selecting students randomly from a national listing will require a smaller sample size, whereas *stratified* and *clustered samples* will require relatively larger sample sizes. Although it may appear contradictory, purely 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.[18]

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 sample clustering, schools are selected first, and then students within schools (clusters) are selected. Picking schools first, and then children, reduces travel costs and travel time and it also 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-morning visit, as a way of increasing sample size at relatively low cost.

Past EGRA applications have shown that it is possible for one enumerator to interview between 12 and 15 children in one school morning.[19] 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, we recommend applying a cluster sampling approach.

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. 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.

To recapitulate, we have discussed four items that need to be included in our sample size calculation. These include:

1. *Variability* in student reading scores (or other EGRA variable if desired)

[18] 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.

[19] 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.

2. Researcher-determined *confidence interval width*
3. Researcher-determined *confidence level*
4. *Design effect (DEFF)* caused by the application of cluster sampling

Calculating sample size for a given confidence interval and confidence level

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

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

where:

n is the sample size needed;
CLtvalue is the t-value associated with the selected confidence level,
DEFT is the square root of the design effect (DEFF), where one uses the DEFT because the squared term gives back the DEFF;
SD is standard deviation, which is a measurement of the variability in our chosen variable;
Width = the researcher-determined *width of the confidence interval;* and
the number 4 is derived from the basic equation for a confidence interval.[20]

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 previous EGRA applications, using the data in Table 1.

The DEFF is calculated as follows:

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

where:

clustersize is the size of the average cluster (the number of children sampled in each school), and
ICC is the intraclass correlation coefficient.

[20] 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.

Increases in *clustersize* or in the ICC will both increase the design effect. If the *clustersize* is 1 (one child per school in the sample), then the ICC does not matter, and the DEFF is 1. That is, clustering does not affect estimated variability if the *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 (our 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 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,

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

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:

$\widetilde{X}$ 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 $MSE_{between}$.

Table 2 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, 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 done thus far. The SD of 29 is a stylized value generalized from the various EGRA studies done thus far.

Table 2. 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 the authors from various EGRA surveys.

The DEFTs in Table 2 above 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.

Recommended Sample Sizes for Confidence Intervals

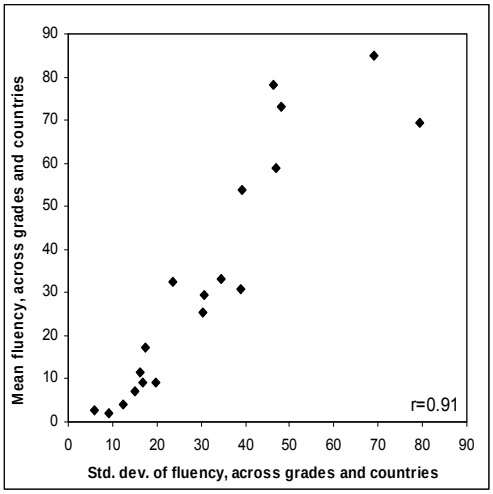
In determining actual recommended sample sizes, a reasonable requirement would be that differences between grades should be "meaningful" in some sense—e.g., the overall confidence intervals should be sufficiently narrow that the confidence intervals for contiguous grades do not overlap. Using Table 1, we can see that the average inter-grade difference is 14. Thus, 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 (as per Table 1 above) the "right" sample size is 409 children, for every grade of interest.

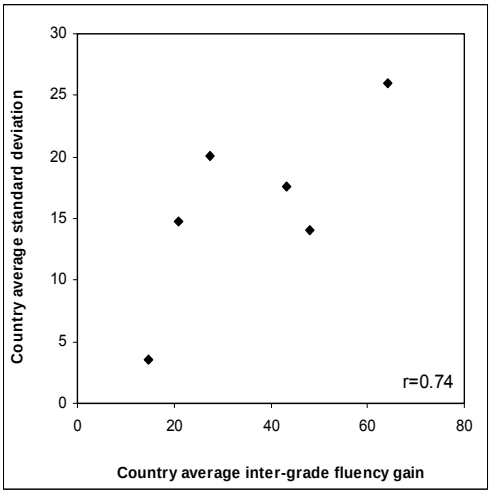
Given the very small differences between boys and girls in the Table 1 sample data (and/or given that the gender differences are highly variable across countries, unlike the steady grade progression), and given the equation for sample size, it should be clear that a very small *Width* would be needed 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 strata of population, 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 thus far appear to be sufficiently small that only very large samples could detect them with statistical significance.

Given that the SDs can be as high as 60 or so, it may seem a bit audacious to propose using an SD of 29 in recommending a sample size. This requires some discussion. First, as can be seen from Table 1 above, the higher SDs tend strongly to be observed only in later grades, and EGRA mostly tries to look at the first two or three grades. It is also the case that SDs appear to be fairly positively correlated with estimated mean fluency levels, and with estimated inter-grade differences (see graphs below).

Graph 1. Mean and SD of Fluency



Graph 2. SD and Inter-grade Gain



In cases where the fluency SDs are high, so are the mean fluency levels, and so are inter-grade differences. This means that in general, differences will tend to be detectable: When the SDs are high, so are the differences one is trying to detect. Or, from a confidence interval point of view, when the SDs are high, so are the central points in the confidence intervals. This means that in principle, a wider *Width* might be sufficient to establish non-overlapping confidence intervals, although a higher sample size is still needed. It is also true that where all that is desired is the confidence interval itself (rather than non-overlap between confidence intervals for different student attributes), the absolute width of the interval matters less than its relative width.

For example, saying that the sample mean is 20, with a confidence interval for the population mean of [10, 30], conveys a lower sense of precision than saying that the sample mean is 45 with a confidence interval for the population mean of, say, [33, 58], even though the latter is

wider in an absolute sense, because the width of the latter is smaller, relative to its midpoint, than the width of the former. Even with these provisos, larger samples are still required if the SD is higher.

Thus, for example, with an SD of 50, but with inter-grade differences now 20 instead of 14 (where 20 is the expected value given an SD of 50, based on Graph 2 above), the sample size would have to go up to 597, as opposed to 409, for each grade of interest.

If one wanted to be more cautious, perhaps sample sizes of 600 students per grade of interest should be recommended. *On the other hand, if with higher SDs one would just decide to be content with a 90% confidence interval, then the required sample size goes back down to 416.* (Many international applications aimed at producing rapid results, such as the World Health Organization's Expanded Programme on Immunization approach in the health sector, use 90% confidence intervals.)

It seems wise to conclude that sampling somewhere between 410 and 600 students per grade of interest (in round numbers) will take care of most contingencies.

Hypothesis Testing Versus Confidence Intervals: Sampling Implications

In deciding about sample sizes, one factor to be taken into account is whether the basis for comparisons between groups (e.g., between fluency levels in different grades) should be 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 only be achieved 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 rather 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 should be used.

Whatever the approach, it is always a good idea to present confidence intervals, and not simply to test hypotheses. Most statistical programs, including Stata, often present both with the same command, so this is not difficult to accomplish. More general-purpose programs such as Excel do not, but the confidence intervals are very easy to generate. The purpose of presenting the CIs is to foster a focus on the parameter in question, such as oral fluency in 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 width of the CIs will tend to be relatively large. Thus, the EGRA approach should decide 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 generalization, that this is due to equal SDs and equal sample sizes.[21] 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:

[21] 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 clarity.

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

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

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

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.

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.

In summary:

If there is a desire to establish non-overlapping confidence intervals, then with parameters similar to what are found in nearly worst-case scenarios, but an SD of 29 (average across those studied), a sample size of 409 students per grade is sufficient.

In cases where the SD is suspected of ranging as high as 50, but where inter-grade differences are correspondingly higher, sample sizes as high as 597 students per grade are needed.

If the aim is to provide one-sided (two-sided) hypothesis tests, then sample sizes of 324 (414 for two-sided) to 472 (604) for two-sided) might be needed for each grade.

For differences by gender, no reasonable sample size is able to detect statistically significant differences, at least judging by the differences by gender observed thus far. The gender differences are just too small.

Summary of Sample Sizes Based on Confidence Intervals and Hypothesis Tests

Table 3 summarizes a range of suggestions on sample sizes. The table assumes an SD of 29, an ICC of 0.45 (which is on the high end of what has been found in EGRA studies thus far), and a *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.

Table 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 the authors.

Sampling Weights and Other Considerations Introduced by Clustering

First summarizing from the discussion above, a simple random sampling approach—such as a simple random sample of children—is difficult to construct and carry out, as very few (none known to the authors) developing countries have national-level rosters of children by individual name. Even if such rosters existed, they would be impractical, as it would be very costly to go assess one child at one school and then travel a great distance to assess another child at another school. Instead, one can first sample schools, and then sample students within schools.

Also as described earlier, rather than sampling 400 students completely at random, which in a large country might mean going to 400 or nearly 400 schools, one first samples, say, 40 schools, and then 10 students per school for the grade(s) of interest. Choosing a fixed number of students per school might keep matters simple (and in some methodologies it may be needed—see section below on Lot Quality Assurance Sampling [LQAS]), but since schools vary

in size, calculating a simple average may be misleading, if larger or smaller schools do systematically better or worse.

As a contrived example to clarify the point: Suppose half of the schools are large (50 students per grade of interest) and half are small (25 students per grade of interest). At the large ones, students have a reading fluency of 20, but at the small ones they have a fluency rate of 50. A *simple* average will yield a fluency of 37.5. A *weighted* average will yield a fluency of 30. This is because the smaller schools, which have the higher fluency, weigh less; the mean is brought down toward the mean of the larger schools. None of these means is inherently "correct." The mean of 37.5 characterizes *schools* and could be used to target schools for intervention; the mean of 30 characterizes *children.*

If the number of children selected per school varies according to school size (say, if one selects every 10th child in the grade of interest), then there is no need to weigh the results, as larger schools will be represented in precise proportion to their size, and the calculated simple mean will represent the learners. In this case, if the researchers wanted an average of schools' reading levels, they would have to calculate the percentage reading at each school, and then calculate the simple mean of those percentages.[22]

One final issue to come back to is that, as noted earlier, if the sampling process uses clusters, and if the fact that clustering was used is ignored when the confidence intervals are calculated, then total variability is underestimated, and the confidence intervals will appear narrower than they really should be.

Several statistical packages are able to take this issue into account. To do so, such packages need to be told what the clustering criterion is (say, schools), and then be told to use a clustering assumption in calculating confidence intervals, and finally be told to calculate using this assumption. For example, in Stata, one declares the clustering criterion to be the school number issuing the command "svyset: school" and then requests a confidence interval on, say, fluency, via the command "svy: mean fluency." To illustrate using countries in which EGRA studies have been carried out thus far, the standard deviations of the mean fluency—assuming no clustering and assuming clustering, respectively—are 3.6 and 4.4 in one case; and 0.57 and 1.31 in another.

Nonparametric Techniques for School-Level Progress Monitoring: A Focus on the LQAS Approach

Most EGRA efforts thus far have been aimed at producing *national-level* awareness that children may not be reading as well as they ought to be, and that the foundation phase in education systems is unduly shaky, making the whole enterprise shaky. These efforts are also aimed at motivating policy makers and implementers to pay attention to this issue and set up remediation processes.

If countries or nongovernmental organizations (NGOs) follow up and attempt remediation and improvement, as some already are doing, then progress monitoring considerations, not just policy awareness, will come to the fore. For broad motivation and analysis, national-level samples of the sizes under discussion in this document are both sufficient and possible, and permit reasonably narrow confidence intervals.

[22] Table 7 at the end of this annex shows an example of these matters.

On the other hand, monitoring over time, usually with repeated measures, at the school level, requires a different approach.[23] EGRA efforts thus far have used a range of from 5 to 20 students per grade per school. School-level samples larger than about 20 per grade are very expensive, particularly if one is going to monitor many schools. Yet, 20 students per grade is not large enough for most traditional statistical estimation of parameters, such as the school-level average reading fluency.

There are two solutions to this problem. First, in a monitoring mode, EGRA could be applied by teachers themselves. That is, at the school level, when teachers do the monitoring, they will assess all children, so the issue of sampling will not enter into the discussion. Having teachers carry out monitoring lowers the cash or fiscal cost (although not necessarily the opportunity or social cost) of monitoring children.

Second, if it is desired to monitor children via an outside agency, then sampling issues do arise, because this approach has cash costs, not just opportunity costs. As noted, sampling more than about 20 children per grade per school, via outsiders, on a monitoring basis, gets extremely expensive. Yet, also as noted, 20 per grade is not a large enough sample to provide much comfort, if one is using the sample to determine average levels of reading in a given school. If one is willing to think nonparametrically, however, sample sizes of 20, or not much more, are good enough to monitor any administrative unit (school or district).

First it may be good to explain the terminology. By a *parametric approach,* we mean an approach based on trying to estimate, say, the average correct words per minute in the school or district (a parameter) via the sample average (an estimator).

There are various problems with the parametric approach, when applied to the school level. One important problem with a parametric approach is the lack of power with such a small sample size. Using the data this document has been manipulating, we note that the cross-country "average of averages" for grade 2 correct words per minute in connected text, in the countries under consideration, is 25, and the SD is 29. The DEFT is not relevant since at the school level one would not cluster. Suppose that with an average at 25, one has a goal of getting to 39 (to carry forth the average inter-grade difference of 14) correct words per minute, i.e., gain one grade, and one needs to detect schools performing at less than that level to know which should receive help.

Even with an SD as high as 29, a hypothesis that "A school with a sample size of 20 is below the target" has a high traditional significance level (*alpha*), with probability value lower than 0.02. However, this test has a power of 0.69 or a *beta* of only around 0.31.[24] This means that while there is a low probability—less than 5%—of helping schools that do not need it (that is, a low probability of accepting the alternative hypothesis when the null hypothesis is true), there is a high probability of not helping schools that do need it (that is, of accepting the null hypothesis when the alternative is true).

Aside from these factors, parametric results might be a little harder to work with, from a monitoring point of view. For an EGRA team interested in monitoring, it might be easier to work with a simple nonparametric rule such as "sample 20 children per grade per school [could be for just one grade—the grade of interest in a monitoring process], and if more than X can read, the

[23] It naturally requires a different approach in terms of the instruments used, and assumes that an intervention has been designed. These issues are important but do not pertain to sampling.
[24] Stata command "sampsi 25 39, sd(29) n(20) onesample onesided."

school does not need help, or does not need the higher level of help." The statement is rendered in absolute numbers, and there are no averages, standard deviations, or percentages to calculate: This is a more typical and simpler monitoring proposition.

Thus, an alternative to parametric statistics is some form of nonparametric approach, where one is not interested in school-level averages, but only in the numbers of children reading at or above a certain level in a given sample of fixed size (or, for that matter, since one may want to monitor processes too, the numbers of children who get exposed to certain instructional techniques). These approaches typically use the binomial distribution to set sample sizes and "decision rules" (such as how many nonreaders can be accepted before it is decided that a school is not meeting goals), in order to keep alpha and beta probabilities down. The most common such technique, thus far mostly used in the health sector in low-income countries, is Lot Quality Assurance Sampling.

To use LQAS, one has to render a given reading variable into a binary variable. The variable takes on the value of 1 if students read above a certain level and 0 if the children are reading below a target correct words-per-minute level. LQAS is frequently used to monitor processes. Thus, alternatively, a value of 1 could be assigned if the students were being exposed to a certain practice and 0 otherwise. The latter does not require turning a continuous variable into a binary variable.

The most common LQAS approach is to think in terms of two risks:

1. the risk (cost) of helping a school that does not need it—usually called "government risk" or "provider risk," as there is a cost in intervening unnecessarily, similar to a type I error in statistics, with associated alpha level; and

2. the risk of not intervening in a school that does need it—usually called "social risk" or "consumer" risk, similar to a type II error in statistics, with an associated beta level.

To minimize provider risk, one wants to set a high level of nonreaders in a school before one helps it. That way one can be quite sure the school needs it. But if one goes too far in this direction, one also runs a risk of *not* helping a school that does need it.

To guard against both of these risks, usually an *upper* threshold is set (for example at 50%), and the number of nonreaders allowed in a sample is set so as to reduce the probability of classifying as nonreading a school that is in fact reading. This argues for tolerating more, rather than fewer, nonreaders in the sample before declaring the school as noncompliant. From the opposite end, a *lower* threshold is determined (for example at 20%), and the number of nonreaders allowed in a sample is set so as to reduce the probability that one would classify the school as not needing help if only 20% or more of children are not reading. This argues for tolerating fewer, rather than more, nonreaders in the sample. In short, one has pressure in both directions. Given those opposing pressures, the final step is to determine the total sample size, and the number of nonreaders to be tolerated before the school is helped, using those thresholds, so as to hold the *total* risk (alpha plus beta) down to some reasonable level, such as less than 15%. In the health sector this level is considered generally appropriate for monitoring purposes.

To repeat, the technique can be used with continuous variables (say, fluency in reading connected text) that have been rendered binary (every child reading at more than X correct

words per minute gets a 1, others get a 0) or with monitoring variables that are naturally binary (such as whether a child is engaged in peer reading every day).

Table 4 illustrates the principles involved, for thresholds of 50% and 20%. The literature from the health sector usually uses examples of 80% and 50% in, for example, monitoring the application of vaccinations. But if one is to use this technique for monitoring reading, it is wise to set fairly low thresholds given how weak the current situation appears to be in most countries studied.

Table 4. Example of LQAS decision table

Sample size	No. of failures, *f*	No of successes, *s*	Cumulative probability of up to and including *s*–1 successes	Cumulative probability of up to and including *f* failures	Total risk
			Equivalent of probability or risk of accepting school as needing help when it does not	Equivalent of probability or risk of not helping a school that needs help	
			Upper threshold: 0.50 (50%)	Lower threshold: 0.20 (20%)	
20	15	5	0.006	0.370	0.376
20	14	6	0.021	0.196	0.216
20	**13**	**7**	**0.058**	**0.087**	**0.144**
20	12	8	0.132	0.032	0.164
20	11	9	0.252	0.010	0.262
20	10	10	0.412	0.003	0.414
20	9	11	0.588	0.001	0.589
20	8	12	0.748	0.000	0.748
20	7	13	0.868	0.000	0.868
20	6	14	0.942	0.000	0.942
20	5	15	0.979	0.000	0.979
20	4	16	0.994	0.000	0.994
20	3	17	0.999	0.000	0.999
20	2	18	1.000	0.000	1.000
20	1	19	1.000	0.000	1.000
20	0	20	1.000	0.000	1.000

Source: Calculated by the authors from assumed values using the binomial distribution.

Reading ***up*** along the 50% threshold column, one can note that to reduce government risk, one wants to accept as not needing stepped-up help only schools one is quite sure have more than 50% readers. One might thus be tempted to accept as not needing help only schools where there are at least 10 readers (since 10/20=50%). But, even schools where as many as 50% of children are reading have as high as a 13.2% chance of producing samples where only up to and including 8 children read, since we are talking only about a sample. A rule based on the expected average will thus be wrong too much of the time.

One might be tempted to reduce this risk further, then, and really accept only as needing help schools with only, say, 6 readers. That would reduce government risk down to 2.1%. Thus, to minimize the risk of accepting too many schools as needing help (thus saving government risk), one could accept lots of nonreaders before accepting the school as needing help. The more

students not reading one is willing to tolerate before deciding the school needs help, the smaller the probability of accepting the school as needing help even though it does not need it—that is, the more sure one can be, based on the sample, that there are enough nonreaders in the school population to tip the school below the 50% threshold. One can see this point by reading up the 50% threshold column. This reduces government risk, or the risk of helping schools that do not need it.

But then (reading across the panel to column 5, the 20% threshold column), going down as low as 6 readers would incur a 19.6% chance of *not* intervening in schools where less than 20% of students are meeting the goal. That is because there is a 19.6% probability, even if only 20% of students are readers, of observing as many as 6 readers or more out of 20. This is too high a level of "consumer risk." We run the risk of not helping schools that need it.

The trick, then, is to add up the two risks, to create a concept of total risk, and look at the pattern of total risk. Thus, with a sample size of 20, the threshold number of readers—that is, the number of readers that minimizes total risk—is 7. The "decision rule" is: Accept a school as not needing help if there are 7 or more readers, or accept as needing help if there are fewer than 7 readers, and this creates a total risk of 0.144. Recall that in a parametric approach, just the second type of risk was as high as 30%. In this case, both alpha and beta risks are kept below 10% and the sum of the two is kept below 15%. If one wanted to drive this risk below 10%, for a 50%–20% set of thresholds, the sample size would have to go up to 26, a fairly high price to pay. Most health applications use 19.

Everything that has been said about children within schools can apply to schools within districts, for example. Similarly, everything that has been said about children reading could be applied to some other school-level practice, such as the school having a reading program of a certain sort. The same sorts of numbers apply.

It would be difficult to explain these rules and ensure that district-level workers in low-income countries could truly understand them, but they are extraordinarily simple to apply on a simple-rules basis, once the basic rules have been set up. It is simply a matter of saying "Take samples of 20. If 7 or more children are reading at a more than X correct words per minute, the school does not get stepped-up help. If less than 7 are, the school needs to be helped more, or put on notice."

The calculation of the probabilities can be programmed very easily. The formula needed in Excel for the first risk (50% threshold) is

=1-BINOMDIST(f,n,1-hithresh, TRUE),

where:

n is the sample size,
f is the number of failures,
hithresh is the higher threshold protecting against government risk, and
TRUE means that it is the cumulative binomial that is desired.

The formula for the second risk is

=BINOMDIST(f,n,1-lothresh,TRUE).

Thus, for example, in order to produce the 2.1% attached to government risk (50% column) above, when more than 6 nonreaders (second row from the top) are tolerated before a school is helped, the researchers would use *1-BINOMDIST(14,20,1-0.5,TRUE*.[25]

As noted, one of the virtues of LQAS is that things can be stated in terms of very simple and fixed decision rules. In Table 5, we have calculated an example that shows the optimal sample size and decision rule (e.g., accept as not needing help a school with 7 or more readers). This can be used by monitors directly, without worrying about all the theory and the more complicated table shown above (Table 4), as is done by health projects.

Table 5. Preset decision rules based on LQAS methodology

Compliance or performance thresholds		For total risk < 0.10		For total risk < 0.15	
Upper	Lower	Optimal sample size	Decision rule	Optimal sample size	Decision rule
0.95	0.65	15	13	13	11
0.90	0.60	20	16	15	12
0.85	0.55	23	17	18	13
0.80	0.50	25	17	20	14
0.75	0.45	27	17	21	13
0.70	0.40	28	16	22	13
0.65	0.35	29	15	23	12
0.60	0.30	29	13	24	12
0.55	0.25	27	11	21	9
0.50	0.20	26	9	20	7
0.45	0.15	23	7	18	6
0.40	0.10	20	5	15	4
0.35	0.05	15	3	13	3

Source: Calculated by the authors from assumed values using the binomial distribution.

Combining and Comparing LQAS and Parameter Estimation

The literature suggests that there is a clever way to combine LQAS and parameter estimation. To do so, school-level data on reading are gathered. But we know that a sample size of even 20 children per grade (or per whole school if one is monitoring only one grade, and noting that in any case, this number is higher than is optimal in a clustered approach) permits neither an evaluation of the percentage of readers as a parameter, nor an estimate of the correct words per minute, also as a parameter, with much confidence.

As noted above, the confidence interval in our most "typical case," with an average correct words per minute of 25, an SD of 29, and a sample size as high as 20 at the individual school level, produces a 95% confidence interval as broad as [11.4, 38.6] at the school level, too broad to work with easily. Accepting a lower confidence level, 90%, does not help much, producing a CI of [13.8, 35.2] at the individual school. Even an estimate of the percentage of readers reading above a certain level produces a very broad CI. In the countries we have studied, a mean of 25

[25] In reality one would use Excel cell references rather than exact numbers.

corresponds to about 20% of learners reading above 40 correct words per minute. A 90% CI for the population percentage is [0.071, 0.401] at school level.

However, the literature suggests that taking samples of 20 students per school, and then perhaps 20 to 40 schools, would make the sample size large enough to allow for parameter estimation at levels higher than the school. This corresponds with the calculations of sample size above that are needed for the *parametric* estimation of either confidence intervals or hypothesis tests, *above* the school level.

This can be quite helpful. One can take an LQAS type of survey to gather baseline parameter data at a level higher than the school. The school-level sample size or cluster size (say, 20), for an initial survey, can be based on what is already known about reading in the country and from evidence around the world. The data can then be used to make nonparametric "yes-no" decisions about *each* school, to establish which schools need more help in the first year of an intervention, and to draw an above-school baseline on the parameters themselves. Repeated measurements can do the same thing, except that one may be able to slightly optimize the within-school sample size once there is a baseline from which to better establish the upper and lower thresholds. Evidently, if the baseline for the school sample size is set at 20, a consideration of DEFT will then drive the total number of schools that should be allowed.

Recommendations for a baseline LQAS sample are shown in Table 6 below. Unlike in the case of standard confidence intervals and hypothesis testing, the within-school sample, or cluster size, is set to 20. That is because this is the lowest that can be accepted for making nonparametric judgments about each school (the school "passes" inspection or not). But then one can use the school-level information, and aggregate it up, to ascertain the (parametric) average level of, say, fluency. The assumption below is that the within-school sample size, or the cluster size, is set to 20.

Table 6. Numbers of schools needed to provide an above-the school parameter estimate, when using LQAS cluster size of 20 at school level

Two-sided confidence width	10	10	15	15
Confidence level	90%	95%	90%	95%
Number of schools needed	45	65	20	28

Source: Calculated by the authors.

In aggregating up LQAS data to estimate parameters, it is important to weight the chosen schools, as was the case above with the standard methods for estimating parameters. Suppose, for example, that the schools in a district were chosen completely at random, using a simple random selection technique. But the schools differ in size. Thus, the 20 students chosen within each school represent different totals. In the contrived example in Table 7, a simple average is shown, and also a weighted average, assuming a small sample of 20 schools and 20 students per school.

Table 7. Using LQAS-sourced data to calculate averages

School no.	Students chosen	Number of readers	Total pupils in tested grade	Weighted readers
1	20	12	50	600
2	20	18	49	882
3	20	7	23	161
4	20	8	25	200
5	20	13	46	598
6	20	11	47	517
7	20	18	48	864
8	20	20	63	1260
9	20	5	17	85
10	20	3	15	45
11	20	6	23	138
12	20	7	19	133
13	20	20	46	920
14	20	19	39	741
15	20	17	48	816
16	20	3	18	54
17	20	4	17	68
18	20	2	26	52
19	20	5	23	115
20	20	8	35	280
Total	400	206	677	8529
Simple mean		10.3		
Weighted mean				12.6

Source: Calculated by the authors.

In this table, the simple mean is calculated by taking the total number of readers (206) and dividing by the total number of schools (20) to get an average number of readers per school of 10.3 (or a reading percentage of 10.3*100/20=51.5%). However, a pattern can be noticed in the data: In each fixed-size cluster or school-sample of 20, larger schools have more readers than small schools do. A simple mean will therefore understate the proportion of children who can read. If one weights the readers at each school according to the number of total students in the grade tested, the reading proportion is much larger. This can be done by multiplying the number of readers at each school times the total number of students in the tested grade at each school, to get the last column. One then divides the total of the last column across all schools by the total in the grade across all schools, to get the weighted average (8529/677=12.6), or a reading percentage of 63%.

The difference between 51.5% and 63% is important, but this is a fairly exaggerated case. In most real-life examples, the difference will not be nearly this large, because there is not likely to be such a strong association between school size and reading ability (0.87 in this contrived example). Also, note that the percentage 51.5% is not necessarily “wrong” and the percentage 63% is not necessarily “right.” The first represents the percentage of pupils who can read in a typical school, and is a valid idea if the unit of analysis is the school: It represents the result of the average school, in terms of how the average school “produces” reading. The second represents the percentage of pupils who can read in the population as a whole, and this is also a valid idea if the unit of analysis is the pupil.

Annex C. 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.

It is strongly recommended that teams directing the EGRA include a person familiar with psychometrics who can run the necessary checks. The below discussion is meant only to offer the reader a brief introduction to the topic and to highlight some of the issues involved. It is not meant to be a comprehensive review; nor does it offer step-by-step instructions for conducting these checks.

Reliability

Reliability may be defined as the degree to which scores for a group of students are consistent over repeated administrations of a test. 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" 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 probably is unreliable.

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 SPSS and Stata can readily compute this coefficient). Cronbach's Alpha may not be the most appropriate measure of the reliability of EGRA scores, however, mainly because portions of the EGRA instrument are timed. Timed or time-limited measures 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.

The **Test-Retest Method** is best suited to estimating the reliability of scores obtained on the EGRA instrument. Test-Retest, which can be conducted as part of the piloting of the EGRA instrument, basically 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 should be 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. The higher the correlation (generally, a value of 0.7 or greater is seen as acceptable), the less susceptible the EGRA scores are to random daily changes in the condition of the test takers or of the testing environment.

A variation on this is to conduct Test-Retest using two similar forms of the EGRA instrument. In this case, the procedure is to administer Form 1 of the test, wait an hour or so, and then administer Form 2. If possible, it is desirable that the order of administration of the forms be reversed for half the group. The correlation (again, a value of 0.7 or above is probably acceptable) between the two sets of scores offers a measure of the degree of stability of EGRA scores over repeated test administrations as well as the degree of equivalence of the scores produced by the two test forms.

Another issue related to the reliability of EGRA scores is **the consistency and accuracy of enumerator performance.** If two enumerators 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? Will either of the enumerators be correct? Since only one enumerator typically listens to and records the responses of each child in the main EGRA administration, the best time to address this consistency (and accuracy) issue is during enumerator training. All enumerators could be asked to listen to the same tape recording of a child taking the EGRA test and to individually record the correctness of the student's responses in terms of the number of correct words read, etc. Any enumerator whose response record has significant errors (e.g., in terms of the discrepancy between the number of correct words per minute recorded by the enumerator and the number of correct words per minute actually read by the student) should receive additional training. If no improvement occurs, they should be removed from the enumerator pool so as not to negatively affect the quality of the data collected during the main study.

Validity

Validity pertains to the appropriateness or correctness of inferences or decisions based on the test results. Returning again to the weighing-scale example, 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 produces results that are not only reliable, but also valid. If the scale consistently reads "20" every time the 30-kg bag is placed on it, then it produces results that are invalid (but still reliable because the measurement, while wrong, is very consistent!).

There is no such thing as a generically valid test. A test's validity must be established with reference to a particular inference or use that is to be based on the test results. Validation is the process by which a test developer or user collects evidence to support the desired inference/use. Validity evidence relevant to the EGRA instrument is described below.

Test-content-related evidence pertains to the degree to which the items on the EGRA test are representative of the construct being measured (i.e., early reading skills in a particular country). The in-country workshop that is held at the start of the EGRA test development process provides an opportunity for countries to build content validity into the instrument by having Ministry officials, curriculum experts, and other relevant groups examine the EGRA template and make judgments about the appropriateness of each item type for measuring the early reading skills of their students. Following this review, these individuals adapt the EGRA instrument as necessary and prepare country-appropriate items for each section of the test. **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 should be collected on 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, namely **evidence of the consequences of test score use** on test takers and

other stakeholders. This 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 D. Open Letter from Deputy Director General, South Africa, to School Principals

AN OPEN LETTER TO ALL PRIMARY SCHOOL PRINCIPALS

Dear Principal

Let's teach our children to read

Reading is a foundational skill that all our children need if they are to succeed in life. Sadly, all our assessments of how well our children read reveal that a shockingly high number cannot read at the appropriate grade and age level. Many simply cannot read at all. We cannot allow this to continue. We are therefore challenging all primary schools to improve the reading skills of all their learners. Even learners who are perceived to be reading at the appropriate level for the grade they are in should be encouraged to move to the next level.

Since the introduction of the National Curriculum Statement, many teachers believe they do not have to teach reading anymore. Nothing could be further from the truth. Reading is probably the single most essential skill a child needs and it should be acquired as early as possible.

There are five areas that are critical for reading:-

- Phonemic awareness (understanding the sounds in spoken words);
- Phonemes (linking sounds to the alphabet and combining these to form words);
- Vocabulary (learning and using new words);
- Fluency (reading with speed, accuracy and understanding); and
- Comprehension (understanding the meaning of what they read).

The Curriculum Guidelines allocate sufficient time for the Language Learning Area and the above should be taught within that time allocation. The Department will provide additional support and guidelines to all schools so principals know how to support teachers.

In addition to the formal teaching of reading, we want all schools to set aside at least 30 minutes a day for the entire school to read (including principal and staff) in any language. This period can be handled in different ways: children who are able to read independently should be encouraged to do so, older learners can read to the younger ones, or teachers to their classes or to groups of children. The Department of Education will help you put storybooks into all classrooms, starting with schools in the poorest communities, so that children can also read for pleasure.

At the beginning of 2006 we put packs of 100 storybooks each, in the different South African languages, into over 5 628 foundation phase classes. Next year we will put similar packs into another 6 000 schools. All we ask is that you help learners read the books and enjoy them.

Later this year we will be testing the reading levels of Grade 3 learners all over the country. A short test to assess whether we are succeeding in our efforts to teach reading will from now on be a regular feature in our monitoring of learner achievement in schools. Take up the challenge and help your children read better!

Palesa T Tyobeka
Deputy Director-General: General Education and Training
Department of Education

education
Department:
Education
REPUBLIC OF SOUTH AFRICA

Annex E. Agenda for Reading Remediation Workshop in Kenya, Based on EGRA Results

DAY 1: Mon **APPROACH**

8:30-9:00 Introductions and setting the goals for the workshop

- Introductions of participants (10 min)
- Overview of workshop goals (20 min) (refer to the above)
  - Review of basic issues surrounding fluency and current thinking about reading more generally
  - Overview of the Early Grade Reading (EGR) project: design and baseline assessment
  - Baseline assessment findings: Overview and implications for remedial interventions
  - Review of remedial interventions: Education for Marginalized Children in Kenya (EMACK) and other-than-Kenya experiences
  - Design of remedial interventions for pre-unit, grade 1, and grade 2
  - Design of learner progress assessment methodologies
  - Testing of designed remedial interventions and improvements
  - Design of the implementation strategy

9:00-10:30 Basic issues around fluency

Why EGR

- Phonemic awareness, phonics, fluency, vocabulary, comprehension

11:00-12:00 EGR project

- Assessment design process, workshop in April:
  - Briefly review the goals of this workshop, participants, and the accomplishments. (30 min)
- Baseline assessment:
  - Overview of implementation: data collection, entry, and analysis (30 min)

12:00-13:00 Purpose of remedial interventions (just a slide or two; more on this topic later)

Presentation of baseline assessment findings, Part I

- Student performance on all tasks in both languages

14:00-16:00 Baseline assessment implications for remedial interventions

- Teacher perspectives on some of the problems encountered in classrooms. We suggest that Aga Khan Foundation (AKF) invite a few teachers (and perhaps Teacher Advisory Centre tutors and Quality Assurance and Standards Officers) to present some of the challenges that teachers face on a daily basis. If AKF selects teacher trainers rather than teachers, they should be teacher trainers who have been teachers in the past 3–4 years.

16:30-17:30 [continued] Baseline assessment implications for remedial interventions

- How to overcome some of those obstacles: involving parents, providing teaching materials, etc.

DAY 2: Tues **APPROACH**

8:30-10:30 Overview of remedial interventions

- Purpose of remedial interventions
- Pratham experience, Peru, South Africa, Guatemala, etc.

Overview of EMACK's teacher training program

- Relevance of EMACK's programs to EGR and possible adoption of some practice

- Review of teacher training delivery mechanisms and implications for EGR

11:00-12:00 [continued] Overview of remedial interventions
12:00-13:00 Pre-unit: Design of remedial interventions (Sylvia Linan-Thompson)
- Analysis of current curriculum
- Identification of components and scope and sequence

14:00-16:00 [continued] Pre-unit: design of remedial interventions
16:30-17:30 Grade 1: Design of remedial interventions
- Analysis of current curriculum
- Identification of components and scope and sequence

DAY 3: Wed APPROACH
8:30- 10:30 [continued] Grade 1: Design of remedial interventions
11:00-12:00 [continued] Grade 1: Design of remedial interventions
12:00-13:00 Grade 2: remedial interventions
- Analysis of current curriculum
- Identification of components and scope and sequence

14:00-16:00 [continued] Grade 2: remedial interventions
16:30-17:30 [continued] Grade 2: remedial interventions

DAY 4: Thu APPROACH
8:30-10:30 Design of learner progress assessment methodologies
- Teacher tools for self-assessment for each grade: pre-unit, 1, and 2
- Tools for EMACK staff: pre-unit, 1, and 2

11:00-12:00 [continued] Design of learner progress assessment methodologies
12:00-13:00 [continued] Design of learner progress assessment methodologies
14:00-16:00 Implementation strategy
16:30-17:30 [continued] Implementation strategy
- Treatment schools:
  - Clarify with AKF what schools will receive treatment. If only treatment schools will be targeted, then we need to make sure that they target the replacement schools and not the ones that were originally selected. If treatment will target all schools in the district, then there is no problem, but then for the post-treatment assessment, we need to select control schools from someplace else.
  - Issue of resources: For instance: If there is only 1 book per 3 children, they need to make sure there are more materials. That might be one reason to focus on just 20 schools: At least we can ensure materials in THOSE schools.
  - Start THIS YEAR with pre-unit and grade 1, next year add grade 2. Next year grade 2's will be tested.
- How to train teachers
- How to provide needed material to them
- How to provide them with the support
- How to organize monitoring

DAY 5: Fri APPROACH
8:30-10:30 [continued] Implementation strategy
11:00-12:00 [continued] Implementation strategy
12:00-13:00 Wrap-up of the workshop

Annex F. Example Lesson Plans for Reading Remediation Based on EGRA Results

Set __1_____ Lesson ___1___ Elements a, s, m, t

Phonological Awareness (5 minutes)	Objective: Given a word, the learner will identify the initial sounds of a word Resources: word list

Word list: as, am, sat, mat, ant, see, man,
Activity:
Tell the students that they will be identifying the first sound in the word you say.
Model the task. The word is "am". The first sound is /a/.
Now let's try it together. The first sound in "am" is… /a/.
Now you try one. The first sound in "as" is …. Students respond.
If students answer correctly, move on to other words.

Phonics (5 minutes)	Objective: Given a letter, name the letter Resources: letter cards: upper- and lowercase a, s, m, t

Letters/letter elements: a, s, m, t
Words:
High-frequency words:
Activity:
Tell the students that they are going to learn the names of some letters. Tell them that after they learn these letters they will begin to read words.
Show each letter and tell students the name. Ask students to repeat the name. After each letter has been introduced, show the letters in random order and ask students to say the name.

Fluency (5 minutes)	Objective: Shown a letter card, the student will name the letter within 3 seconds. Resources: Letter cards for upper- and lowercase a, m, s, t

Focus: Letters
Activity: Show students each letter and ask them to name it. If they do not name it within 3 seconds, tell them the name and ask them to repeat it. Continue with other letters.

Vocabulary and comprehension (15 minutes)	Objective: TLW listen to a sentence and identify who the sentence is about. Resources: Sentence written on board or chart.

Activity:
Tell students that you will be reading a sentence.
I am Sam.
Who is the sentence about? Sam.