March 2014

Early Grade Mathematics Assessment (EGMA) Toolkit

RTI INTERNATIONAL

Acknowledgments

This toolkit is the product of the collaboration of many different members of the international mathematics education community, and was entirely funded by RTI's strategic investment funding.

We would like to thank the children, teachers, school officials, assessors, and community members in participating countries. Their support is at the heart of this document, and we appreciate their willingness and contributions.

The authorship of this toolkit was a collaborative effort between Linda M. Platas, Leanne Ketterlin-Gellar, Aarnout Brombacher, and Yasmin Sitabkhan.

Table of Contents

SECTION PAGE

List of Figures

NUMBER PAGE

List of Tables

NUMBER PAGE

Abbreviations

ALPP	Accelerated Learning Program PLUS
CESLY	Core Education Skills for Liberian Youth (USAID project)
DIF	differential item functioning
EdData II	Education Data for Decision Making (USAID project)
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
MANOVA	multivariate analysis of variance
NCTM	National Council of Teachers of Mathematics
TIMSS	Trends in International Mathematics and Science Study
USAID	United States Agency for International Development

Definitions

Domain: In the fields of education and development, domain refers to an area of knowledge. In this toolkit, this term generally refers to the domain of mathematics.

Subdomain: A subset of an area of knowledge from a specific domain. In this toolkit, this term generally refers to areas of knowledge within mathematics, such as counting, geometry, or arithmetic.

Subtest: In this toolkit, this term refers to one of eight short tests within the Core Early Grades Mathematics Assessment. In the EGMA toolkit, the subtest are referred to as tasks. Each short test assesses a unique aspect of numeracy. These include *Number Identification, Number Discrimination, Missing Number, Addition Level 1, Addition Level 2, Subtraction Level 1, Subtraction Level 2,* and *Word Problems.*

Numeracy: Used in this toolkit to reference the trait of having number sense. Includes fluency and flexibility with numbers, use of numbers to describe both real and abstract entities, and the ability to perform mental mathematics.

Numerosity: The quantitative quality of a set (e.g., four dogs, six blocks, two numbers).

Informal Mathematical Skills: Used in this toolkit to refer to the mathematics knowledge and skills learned prior to and/or outside of formal schooling. It includes mathematics learned through a variety of experiences such as (but not limited to) trade, game-playing, and currency transactions.

Formal Mathematical Skills: Used in this toolkit to refers to mathematics knowledge and skills learned in school, and includes symbolic representation.

Mathematical Literacy: Used in this toolkit to refer to the ability to use quantitative skills in everyday life (making purchases, paying bills, borrowing money, etc.) and understand information that is presented in mathematical formats (graphs, percentages, ratios, etc.)

Chapter 1: Introduction to the Core EGMA

A strong foundation in mathematics during the early grades is the key to future success in mathematics, which is instrumental in the development of workplace skills and knowledge (Malloy, 2008; Nunes & Bryant, 1996; Steen, 2001; U.S. Department of Education, 2008). In addition, basic mathematical reasoning is key to everyday activities such as shopping and personal finance. Recent meta-analyses also suggest that early mathematics skills predict later reading skills just as much as early reading skills (Duncan et al., 2007; Romano et al., 2010). There is a growing recognition among policy makers, donors, and educators, of the importance of making evidence-based policy and programming decisions. However, Ministries of Education in developing countries and donor organizations are challenged by a lack of solid information about student learning outcomes in mathematics, particularly in the early grades. Building on the success of, and great demand for, the Early Grade Reading Assessment (EGRA), which was developed in 2006 and implemented to date in more than 50 countries, the United States Agency for International Development (USAID) contracted with RTI International in 2008 to develop an assessment of early grade mathematics competencies. The result was the Early Grade Mathematics Assessment (EGMA), an orally administered assessment of the core mathematical competencies taught in primary grades[1]. The EGMA, to date, has been used in 14 countries around the world.[2] As is noted later in this and subsequent chapters, this original EGMA was revised in 2011 and renamed the Core EGMA. This Toolkit provides information on the use of the Core EGMA.

Data from large, cross-national mathematics studies that have included developing countries (e.g., the Trends in International Mathematics and Science Study [TIMSS]) have demonstrated that those countries lag behind the developed countries in mathematics performance. These studies do not, however, offer information about what is causing this poor performance or where it starts. The Core EGMA, in contrast, offers an opportunity to determine whether children are developing the fundamental skills upon which other mathematical skills build, and, if not, where efforts might be best directed. This is vital information for countries that are working to improve the quality of education in their schools.

1.1 What Is the Core EGMA?

The Core EGMA is an assessment of early mathematics learning, with an emphasis on number and operations. The Core EGMA consists of six subtests (referred to as *tasks* in the instrument) that, taken together, can produce a snapshot of children's knowledge of the competencies that are fundamental in early grade mathematics. These competencies (and their respective subtest titles) include number identification (Number Identification), reasoning about magnitude (Number Discrimination), recognition of number patterns (Missing Number), addition and subtraction (Addition and Subtraction, Levels 1 and 2), and word problems (Word Problems). The Core EGMA is an oral assessment and individually administered to students by trained assessors. Many mathematics assessments require children to be able to read in order to solve problems. Because the Core EGMA is designed for the early grades, which is when children are just beginning to learn how to read, the oral administration does not confound a child's ability to read or write with a child's ability to do mathematics.

[1] The Core EGMA was designed to assess children's early mathematical skills in primary grades first through third.

[2] The EGMA has been administered in the Democratic Republic of Congo, Dominican Republic, Ghana, Iraq, Jordan, Kenya, Liberia, Malawi, Mali, Morocco, Nicaragua, Nigeria, Rwanda, and Zambia.

The Core EGMA items were drawn from extensive research on early mathematics learning and assessment and were constructed by a panel of experts on mathematics education and cognition (see panel details in Section 2.6). The conceptual framework for mathematical development is grounded in extensive research that has been conducted over the past 60 years (e.g., Baroody, Lai, & Mix, 2006; Chard et al., 2005; Clements & Sarama, 2007; Ginsburg, Klein, & Starkey, 1998). Because the Core EGMA is designed to guide instructional decisions, it is important for the assessment to sample skills that may indicate the need for intervention (Fuchs & Fuchs, 2004). These early numeracy skills assessed by the Core EGMA are key in building students' abilities to solve more advanced problems and to acquire more complex mathematics skills (Baroody et al., 2006; Clements & Sarama, 2007). Additionally, the Core EGMA should be quick to administer, have high face validity, allow for alternate forms for multiple administrations, and result in reliable and valid data (Clarke, Baker, Smolkowski, & Chard, 2008).

The Core EGMA, much like the EGRA, is meant to be locally adapted to fit the needs of the local context. Although item specifications have been developed to preserve the mathematical integrity of the subdomains that are being assessed, there are significant places in which the assessment must be adapted to the local context; not the least of these is language. Although many view language as being unrelated to mathematics, in reality, language is integral to the use and learning of mathematics, and it exemplifies many of the cultural aspects of mathematics (Gay & Cole, 1967; Klibanoff, Levine, Huttenlocher, Vasilyeva, & Hedges, 2006; Nunes & Bryant, 1996; Saxe, 1991).

1.2 Why Is the Core EGMA Important?

Although other assessments measure mathematical knowledge, the Core EGMA specifically focuses on the skills in the early grades of primary school. In the early grades, there is a strong emphasis on numeracy, which is often called "number sense." Numeracy refers to a "child's fluidity and flexibility with numbers, the sense of what numbers mean, and an ability to perform mental mathematics and to look at the world and make comparisons" (Gersten & Chard, 1999, p. 19–20). The Core EGMA focuses primarily on numeracy. This is important because the Core EGMA identifies gaps in the mathematics education that children are receiving at an early age. To address this skill differential, students with limited "number sense" may need to be identified early and provided with targeted interventions that provide extra support in learning fundamental mathematical concepts.

In developed countries, interventions might be applied to individual children or small groups; however, in developing countries, where large numbers of students may be underperforming, interventions must be geared toward entire systems. Assessments of students' mathematical knowledge and skills are thus needed to determine the level of support needed, whether at the national, district, or local level. The Core EGMA takes a step toward addressing this need.

1.3 What Are the Uses of the Core EGMA?

Assessments usually come in two forms: summative and formative. Summative assessments are generally thought of as occurring at the end of a period of instruction and are designed to determine whether particular academic goals or benchmarks were reached. Generally, there are no opportunities to teach the content again following summative assessments. The TIMSS is such an instrument and, in fact, is described as an "international comparative study of student achievement" (Provasnik et al., 2012, p. iii). Summative assessments can also be used to measure learning over time, comparing different curricula or interventions. In these instances, the assessment is used to determine students' achievement at the beginning and end of a specified period of time, with the goal of determining whether a curriculum or intervention was more successful than "business as usual" or perhaps another curriculum or intervention.

In contrast to summative assessments, formative assessments are designed to create opportunities to revise instruction to provide more support for students where they need it. Generally, formative assessments are used by teachers or schools to inform teaching practices and more closely tailor instruction to students' location in a developmental progression. As an example, a unit in a curriculum will be taught in a classroom, and then a formative assessment will be conducted of the knowledge expected to have been gained during that unit. Depending on the outcomes of the formative assessment, a teacher may teach specific concepts again to the whole class, work with individual students to bring their knowledge and skills up to the majority of students in the class, or move on to the next unit.

The labels "summative" and "formative" do not necessarily dictate the format of the assessment. For example, a state, district, province, or other entity may want to understand, at the end of the school year, what percentage of fifth-grade students have a conceptual understanding of addition of fractions. Similarly, a teacher may want to know whether most of his or her students have a conceptual understanding of addition of fractions before he or she moves onto the next content area in the curriculum. Both types of assessments may include items asking students to solve problems involving the addition of fractions and provide explanations of how they solved the problems. In short: Same problem, different purposes.

The Core EGMA holds promise for both summative and formative assessments. The Core EGMA was originally designed to provide both types of information. However, for the most part, the Core EGMA has been used to (1) determine how students in a country are performing overall compared to its stated curriculum and (2) examine the effectiveness of specific curricula, interventions, or teacher training programs. The following paragraphs describe two possible types of uses for the Core EGMA: (1) as a country-level diagnostic tool (Section 1.3.1) and (2) for program evaluation (Section 1.3.2). Section 1.4 describes types of uses for the Core EGMA that are discouraged.

1.3.1 Country-Level Diagnostic Tool

An instrument such as the Core EGMA holds significant promise in providing stakeholders, from Ministries of Education to aid agencies, with the information essential to making informed local and national system changes in teacher education and curriculum development and implementation. Accurate and efficient assessment of early grade mathematical skills does not provide all of the information necessary to effect change for the better in mathematics teaching and learning. However, adding an instrument that is valid across national borders and in multiple languages can provide the Ministries of Education with an overall assessment of students' abilities based on what is known about children's mathematical development, independent of regional curricula. Depending on the level of analyses, this information can be tailored to examine specific within-country geographical or political sub-regions. The results of these analyses can give Ministries of Education the ability to target interventions, such as teacher training, on specific topics or sub-regions. It is important to understand that students' scores are more than just a measure of their abilities in a specific skill—they are also a measure of the students' environment. Myriad factors can affect student learning, including socioeconomic status, nutrition, skills and education of teachers, instructional time, parental education, civil war or localized violence, and family responsibilities.

1.3.2 Program Evaluation

So far, the most common use of the Core EGMA has been to measure change over time, usually during the course of a new curriculum, teacher training program, or intervention. As is known from many studies worldwide (Martin, Mullis, & Foy, 2008), the use of a valid and reliable measurement instrument is crucial during such evaluations. Such an instrument allows programmatic evaluations that are not based on benchmarks specific to any one curriculum or intervention, and identifies strengths and weaknesses in what we know are the foundational to higher-level mathematics.

Classroom-Level Formative Assessment

The Core EGMA could be used as a formative assessment tool to provide diagnostic feedback that could be made available to teachers and schools. As stated in the report titled *Early Grade Mathematics Assessment (EGMA): A Conceptual Framework Based on Mathematics Skills Development in Children* (RTI International, 2009a), one of the objectives of the original EGMA was to choose, and present, the measures in such a way that teachers could see how they relate to the curriculum. Although many curricula in developed and developing countries follow what is known about children's progression in their mathematical development, there will most likely not be a direct match between curricula and the Core EGMA. However, because the Core EGMA taps into foundational skills, upon which later and more complex mathematical thinking is based, it could be highly useful to employ the Core EGMA at the beginning of the year to determine where children are in their mathematical development for the purposes of informing instructional practices.

1.4 How Should the Core EGMA Not Be Used? (Uses and Interpretations of Results)

The following paragraphs describe four scenarios for which the Core EGMA should ***not*** be used: (1) for cross-country comparisons, (2) for high-stakes testing, (3) as input for student report cards, and (4) for simultaneous program and country-level diagnostics.

1.4.1 Cross-Country Comparisons

The Core EGMA was not designed to compare students' mathematical development across countries. This is especially true for developing countries, whose learning environments vary widely. For example, some countries are further along in their efforts to offer universal primary school access. Even so, those that have been able to broaden access may have difficulty with providing high-quality instruction due to an insufficient supply of qualified teachers. Cross-county comparisons do not provide information on how to reduce differences in achievement, and thus serve no purpose. However, in the future, it may be useful to develop cut-points in the Core EGMA to give Ministries of Education more information on which grade levels may benefit from interventions.

1.4.2 High-Stakes Testing

The Core EGMA was not designed to support high-stakes testing of students or evaluations of teachers. As previously mentioned, the Core EGMA is purposefully not tied to the curriculum of any one country. The Core EGMA does not explain the "how" or "why" of students' or teachers' performance; therefore, the instrument does not yield sufficient information to make decisions about funding or educational or career opportunities.

1.4.3 Direct Scoring of Students by Teachers for Report Cards

The purpose of report cards is to supply parents with information about students' progression through a specific program or curriculum and to indicate whether they are meeting grade-level expectations. The Core EGMA is not a program or curriculum and does not directly assess either.

1.4.4 Simultaneous Program Evaluation and Country-Level Diagnostics

Program evaluation describes a process in which a specific program, curriculum, or intervention is examined. Country-level diagnostics measure students' performance at a point in time to provide Ministries of Education with a snapshot of where students are in the assessed subject area (e.g., reading or mathematics). In other words, program evaluation measures how effective a specific program is, and country-level diagnostics evaluate how well the entire system is working.

In the case of the Core EGMA, it can measure how effective a program is at providing effective instruction in the support of foundational mathematical skills. Alternatively, the Core EGMA can measure how well a country-wide system is supporting students' learning in foundational mathematics. However, because program evaluations only measure implementation of a specific program in specific locations, and country-level diagnostics measure country-wide systems regardless of program or location, the Core EGMA cannot measure both simultaneously.

1.5 Outline of this Toolkit

The remainder of the toolkit provides details about three topics essential to understanding effective implementation of the Core EGMA. Chapter 2 presents details of the Core EGMA's development, including the theoretical and research bases of the mathematical concepts being measured, theories of measurement and how the Core EGMA was developed using the best practices identified in the field. Chapter 3 discusses the technical adequacy of the Core EGMA and details the validity and reliability evidence collected. Chapter 4 specifies how to adapt the Core EGMA to local contexts, describes how to train assessors to collect data, and offers recommendations for data analysis.

Chapter 2: Development of the Core EGMA

2.1 Theoretical Perspective of the Core EGMA

The acquisition of mathematical knowledge begins with early competencies upon which children begin to construct more complex understandings, which are informed by sociocultural experiences (Saxe, 1991). The development of these skills begins long before formal schooling (Ginsburg & Baron, 1993; Sarama & Clements, 2009). Competencies that are innate or that develop in the first years of life, such as attending to quantities or spatial orientation, provide foundations for later environment-dependent competencies. The study of early mathematics has a rich and varied history, with each theorist or researcher building on or reacting to earlier research (Ginsburg et al., 1998). Current understanding of how mathematical development occurs is informed by these successive and simultaneous efforts.

The Core EGMA is built on the assumption that prior to formal schooling, children possess informal mathematical skills – those mathematical skills gained outside a school context[3]. All societies use mathematical knowledge to some extent. This knowledge includes the use of mathematics to keep track of things or people, engage in trade or selling activities, and describe direction or distance and measurements (Radford, 2008). Spatial knowledge is used to remember locations and distance, understand part–whole linkages (e.g., an object can be visualized that is partly occluded by a tree), and classify knowledge to help organize environments. Children construct knowledge through actions in their environments, such as finding appropriate stones or seeds for *mancala* games, gathering foodstuffs, dividing a set of toy trains among siblings, or playing with blocks.

Many of the previously mentioned abilities require few or no language skills. However, much of mathematics requires a special vocabulary that differs from language to language (and even dialect to dialect). Because of the vocabulary differences and because mathematics is a culturally constructed domain with broadly accepted conventions, mathematical development is also dependent on sociocultural interactions (Gay & Cole, 1967; Nunes & Bryant, 1996; Saxe, 1991; Vygotsky, 1978). Most societies have developed a complex mathematical vocabulary, including names for numbers and shapes, terms for spatial orientation, names for numerical operations such as *division* and *addition*, and mathematical terms such as *fraction*, *integer*, and *factor*. Thus, young children learn number names and mathematical terminology because peers or adults use them in interactions, not through their own solitary actions (Perry & Dockett, 2008).

Formal schooling generally includes introducing symbolic representation in mathematics (Langrall, Mooney, Nisbet, & Jones, 2008). Prior to schooling, a child may be able to count four objects in a set and pronounce that there are "four." Formal schooling provides instruction that assists the child in connecting the numeral "4" with such a set. Young children know that a set of four biscuits can be evenly divided between two friends, but it is not until school that they learn the accompanying equation, $4 \div 2 = 2$. In addition to teaching how to use written notational systems, formal schooling provides instruction in codified mathematics. As previously mentioned, formal mathematics is a cultural construction, built over millennia. Although algorithms, formulas, axioms, and functions are primarily based on real-world applications, they provide a means to convey the abstract aspects of numbers—e.g., *four* is an attribute of entities, not an entity in and of itself—and therefore can apply to discrete quantities, measurement,

[3] Informal mathematics generally refers to mathematics learned prior to and/or outside of formal schooling It includes mathematics learned through a variety of experiences such as trade (e.g., tailors and carpenters), game-playing (e.g., board games or scorekeeping in football or basketball), and currency transactions (e.g., street sellers, and market purchases). Formal mathematics refers to mathematics learned in school.

temperature, strength, or force. Put simply, mathematics enables us to investigate, organize, understand, and describe the world around us.

2.2 Importance of Mathematical Literacy

The preceding paragraphs briefly discussed the acquisition of mathematical knowledge in both preschool and early primary settings. Although knowing this overall progression is important, it tells us little about the value of learning formal mathematics or how we might go about understanding where children are in the developmental progression.

To address the first point—acquisition of mathematical knowledge in both preschool and early primary settings—mathematical literacy has increasingly become a required skill in everyday life (Steen, 2001). Citizenship requires numeracy. For example, deciding which candidate in an election can provide the best opportunities for the population requires interpretation of outcomes such as inflation, debt, and export and import ratios. In addition, voting for or against tax increases requires an understanding of the costs and benefits to such taxes. Many, if not most, occupations require numeracy. Examples include not only science and technology occupations, but also those in law, architecture, and food preparation. In addition, personal finance, whether one possesses very little or very much money, requires numeracy to judiciously manage it. Microfinance loans have caught many industrious individuals unaware because of a lack of understanding of interest rates. Another area that requires numeracy involves health maintenance. For example, determining the risks of malaria requires a rudimentary understanding of statistics to answer some key questions such as the following: How effective are mosquito nets? Should they be treated with insecticide? What type of malaria is endemic in any one area? What drugs are effective for that species?

With regard to developing countries, the argument is sometimes made that many children will grow up to be subsistence farmers, herders, or gatherers, so why waste valuable state resources on these groups if they will use only the most basic of mathematic skills? However, examples of the importance of quantitative literacy abound, even for children in remote and isolated villages (Steen, 2001). In addition to the previously mentioned reasons for ensuring mathematical literacy, the most important reason for providing effective mathematics instruction to *all* children is that without such instruction, future paths to opportunities may be permanently closed. Formal mathematics allows us to abstract quantities and linkages, apply them to new contexts and situations, and generalize instead of having specific skills only in one area or tied to one context.

2.3 Measurement of Mathematical Literacy

How do we measure whether children are achieving mathematical literacy? More to the point, how do we measure whether educational systems at national or subnational levels are creating opportunities for students to achieve mathematical literacy? Although these questions appear to have similar goals, a major difference is that one measures individual children's progress along a trajectory, and the other measures whether educational systems produce students with the mathematical skills necessary to prosper in the twenty-first century. The Core EGMA was designed primarily to measure performance to answer the latter question[4]. The Core EGMA (details in Section 2.6.5) was designed to assess, on a large scale, the extent to which students have gained specific foundational mathematical skills. By "foundational," we do not mean what exists in a country-specific curriculum in the early years

[4] As noted in Chapter 1, the Core EGMA could also be used as a formative assessment, informing teachers where children generally are in the assessed skills at the beginning and end of the school year. However, as the Core EGMA is not tied to any one curriculum, its utility as a formative assessment tool throughout the school year is limited.

of school, but what research says is reasonable, given what is known about cross-country achievement in mathematics. These skills (see also Section 2.4) include number identification, number discrimination, recognition of number patterns, and addition and subtraction with one- and two-digit numbers.

The development of valid and reliable assessments of mathematics requires an understanding of which skills are to be measured and how these skills can be accurately measured. The National Council of Teachers of Mathematics' *Principles and Standards for School Mathematics* (NCTM, 2000) states that instruction should enable students to "understand numbers, ways of representing numbers, relationships among numbers, and number systems." This statement is in accord with those topics deemed worthy of assessment in the TIMSS (Booth & Siegler, 2008, p. 4), which assesses mathematical knowledge based on "what society would like to see taught (the intended curriculum), what is actually taught (the implemented curriculum), and what the students learn (the attained curriculum)."

Validity and reliability of the Core EGMA are discussed in Chapter 3 of this toolkit. However, a short discussion on the linkage between the development of the Core EGMA and the types of validity required to support its use is in order. Validity can be defined by the extent to which evidence and theory support the interpretation of assessment outcomes (Organisation for Economic Co-operation and Development [OECD], 2009). The validity of an instrument can be measured in many ways. Of these various ways, content validity, predictive validity, and convergent and discriminant validity most pertain to the Core EGMA (Sarama & Clements, 2009). An important aspect of validity, that of the consequences of test interpretation (De Smedt, Verschaffel, & Ghesquiere, 2009), particularly and additionally pertains to the use of the Core EGMA.

The validity of the Core EGMA is directly related to its design and utility. As discussed in the following paragraphs, the Core EGMA is designed to predict later achievement (predictive validity) and align with curricula that support the development of foundational mathematical skills (content validity). The Core EGMA is also designed to identify the linkages among those skills (convergent validity), to discriminate among types of mathematical knowledge, and to differentiate among skill levels.

Similarly to the TIMSS, the Core EGMA is designed to assess mathematical knowledge that the intended curriculum is supposed to engender. For the Core EGMA, however, this knowledge consists of foundational mathematical skills upon which later, more sophisticated mathematical understandings are built (e.g., those measured by the TIMSS in fourth grade and later). The Core EGMA is also designed to assess students' learning of such a foundational curriculum, which is ultimately tied to what is actually taught in classrooms. All of these are connected: (1) the state creates a curriculum based on what is known about foundational mathematics skills, (2) teachers seek to enact instruction that provides opportunities for students to learn these skills, and (3) if done effectively, students progressively gain more mathematical skills that can be measured.

To be maximally useful to an educational system (e.g., national or subnational), an early mathematics assessment should assess those skills that are ***foundational to later and more sophisticated mathematical abilities; predictive of later mathematical achievement; and teachable.*** (Note that the subtest descriptions in Section 2.4.1–2.4.5 are explicitly linked to these requirements.) In the best of all worlds, the curriculum enacted in the schools in which the students are being assessed should highly correlate to those foundational and predictive skills. Finally, a useful early mathematical assessment must be able to measure progress in these skills.

An important aspect of validity is the elimination of possible confounds in the interpretation of outcomes. Some examples of the possible confounds include the language in which the assessment is conducted (low scores might reflect a lack of language-of-assessment skills rather than mathematical skills) and the symbols used in the assessment (low scores might reflect a misunderstanding of the

symbols rather than a lack of the underlying mathematics skills). An additional example is the method of administration (low scores on paper-and-pencil tests might reflect a lack of experience with writing utensils or the inability to write, rather than a lack of mathematics skills). To reduce these possible confounds, the Core EGMA was designed to be administered in the language in which the student is most mathematically facile. The Core EGMA instrument adaptation before each administration ensures that it reflects the set of mathematical symbols most common within the educational system. The Core EGMA is also an oral assessment, so no student writing is required. Answers to all items are orally transmitted from a student to an assessor.

Fatigue is another related confound. Fatigue can affect outcomes on items toward the end of an assessment due to inattention, boredom, anxiety, and cumulative cognitive load. Two attributes of the Core EGMA are in place to help prevent fatigue: (1) the length of the assessment and (2) the existence of stop rules. Regarding the first attribute, for assessments to be useful, they must be short enough so that students are not unduly taxed (Piazza, Pica, Izard, Spelke, & Dehaene, 2013), but long enough so that accurate information is obtained. Additionally, because the Core EGMA is generally used to measure students within educational systems, it must be short enough to be administered one by one to a sample of students in a single school on the same day. The approximate administration length of the Core EGMA is 20 minutes. This length of time generally prevents test fatigue, yet is long enough to allow for a sufficient number of items to accurately measure each subdomain. Regarding the second attribute, the use of stop rules helps to prevent student frustration and fatigue. Students in testing environments are already stressed, and the added stress of repeated struggles to provide answers to questions they do not know creates an even more undesirable testing environment. Stop rules enable an assessor to stop the administration of a subtest when the student has incorrectly answered a specific number of items in a row (generally four). Stop rules can also be used when students do not provide answers for a specific number of items in a row.

2.4 Mathematical Subdomains of the Core EGMA

The Core EGMA measures foundational mathematical skills. Unlike literacy, mathematical skills continue to build upon each other throughout a lifetime. That is, although vocabulary and fluency increase over a lifetime, this is different from an understanding of probability building on an understanding of operations, for example. Unlike symbols used in reading (alphabetic-, syllabic-, or word-based), lower numbers and numerosities are learned first. For example, the letter “F” is not intrinsically more difficult to learn than the letter “B,” but a set of six objects is more difficult to count than a set of two objects. The consequences of this slow-building understanding of more complex mathematical understandings are that specific mathematical concepts are foundational to others. As mentioned earlier, these specific mathematical subdomains include

- number identification,
- number discrimination (which numeral represents a numerosity greater than another),
- number pattern identification (a precursor to algebra), and
- addition and subtraction (including word problems).

Each of these is elaborated in the subsections that follow. Section 2.5 notes other important subdomains that were omitted from the Core EGMA to limit its length.

In addition to the goal of measuring foundational mathematics, alignment with common curricula in mathematics is also a self-imposed requirement of the Core EGMA. In a review of countries participating in the TIMSS assessment, Schmidt and colleagues (2005, p. 534) listed whole-number meaning and operations as the top two content areas in grades 1–3 for all of the A+ TIMSS participating countries. Along with rationales for the Core EGMA content, content validity is also discussed in this

section. Content validity is based on the common acceptance by experts of a set of items that represent a known construct or concept. In literacy (specifically English), the alphabet is the basis for the concept of alphabetic knowledge. In numeracy, knowledge of numerals and an understanding of their numerosity are the basis for the number competency.

Research, in particular since the middle of the last century, has supported the idea argued in this section of developmental progressions in early mathematics (Baroody, 2004; Clements & Sarama, 2007; Ginsburg et al., 1998). Within subdomains, this progression is usually linear, from problems involving smaller numerosities to those with larger numerosities. However, there is not always a developmental progression from subdomain to subdomain. So, although division may be considered a more sophisticated skill than addition, young children may learn how to divide sets of objects (the basis of fair-sharing activities with food or toys, for instance) at the same time that they are learning addition and subtraction (e.g., "If you take one more from me, you will have more than me"). In spite of this lack of consistent linear developmental progressions between subdomains, curricular progressions are highly similar in early mathematics. The arrangement of the subtests within the Core EGMA—also reflected in the subsections that follow—takes into account this widely accepted concept of progression from lower-order to higher-order skills."

The following section describes the subtests of the Core EGMA. As noted earlier, all subtests meet the three requirements for inclusion: They reflect the curriculum, are predictive, and are teachable.

2.4.1 Number Identification

Description: The Number Identification subtest is timed (60 seconds) with no stop rules, and it consists of 20 items that increase in difficulty. The first three items of the subtest include the numerals 0, 9, and one other single-digit number. The next 12 items consist of two-digit numbers from 10 to 99, and the last five items are three-digit numbers from 100 to 999. Students are asked to say each number aloud.

Number Identification items

2	9	0	12	30
22	45	39	23	48
91	33	74	87	65
108	245	587	731	989

Background: Number competence is reflected in counting procedures, fact retrieval, and accurate computation (Jordan, Kaplan, Ramineni, & Locuniak, 2009). At the base of these competencies is the ability to identify numerals and their numerosities, which is why number identification is frequently the first mathematics topic taught in the early years of primary school (Geary, 2000). As indicated earlier in the toolkit, formal schooling brings students into contact with the abstract symbols that represent quantities and operations. Unlike pictographs, in which a graphic resembles a physical object (e.g., a drawing of a star that depicts a star or aspects of the Mixtec language in Mexico), numerals do not resemble the numerosities that they represent. Therefore, students must commit to memory the numeral, number word, and numerosity/magnitude. Although the ability to count orally and establish one-to-one correspondence is at the basis of all later mathematical abilities (Gelman & Gallistel, 1978), it is the subsequent comprehension of the numerosities that numerals represent that is at the basis of abstract and symbol-based mathematics.

In the Hindu-Arabic number system, the numerals include 0 through 9. Singly or combined, these characters can represent any number in the counting system. The system also uses position and decimal points to represent these numbers (Baroody, 1987a). Because the position of these numerals makes a difference in the interpretation of the numerosity, understanding place value (e.g., the value of a "9" in the ones and tens place means 9 and 90, respectively) is essential in conceptual understanding of number values. Therefore, Number Identification consists of both single- and multi-digit items, and

correct answers must reflect the place value in multi-digit numerals (e.g., nine-nine is not a correct answer to 99, but ninety-nine is).

2.4.2 Number Discrimination

Description: The Number Discrimination subtest is an untimed test of 10 items with a stop rule after four successive errors. Each item consists of a set of two numbers, one of which is greater than the other. The first item is a set of one-digit numbers, the next five items are sets of two-digit numbers, and the last four items are three-digit numbers. Students state the higher of each set of two numbers (pointing at the correct number is insufficient evidence for scoring).

Number Discrimination items

7	5	94	78
11	24	146	153
39	23	287	534
58	49	623	632
65	67	867	965

Background: Performance on comparisons of numerical magnitude, such as comparing the number 23 with the number 32, are predictive of later mathematical achievement (De Smedt et al., 2009). A theoretical reason for the predictability is that an understanding of numerical magnitudes can narrow the number of possible answers in solving arithmetic problems (Booth & Siegler, 2008). Another theoretical reason for is that early arithmetic problem solving frequently evolves to the "min" procedure (i.e., counting on from the highest number; Geary, Bow-Thomas, & Yao, 1992). Brain imaging also supports the idea that the area of the brain responsible for processing of magnitude is active during arithmetical tasks (Dehaene, Piazza, Pinel, & Cohen, 2003). Finally, an understanding of magnitudes can allow children to determine the plausibility of their answers.

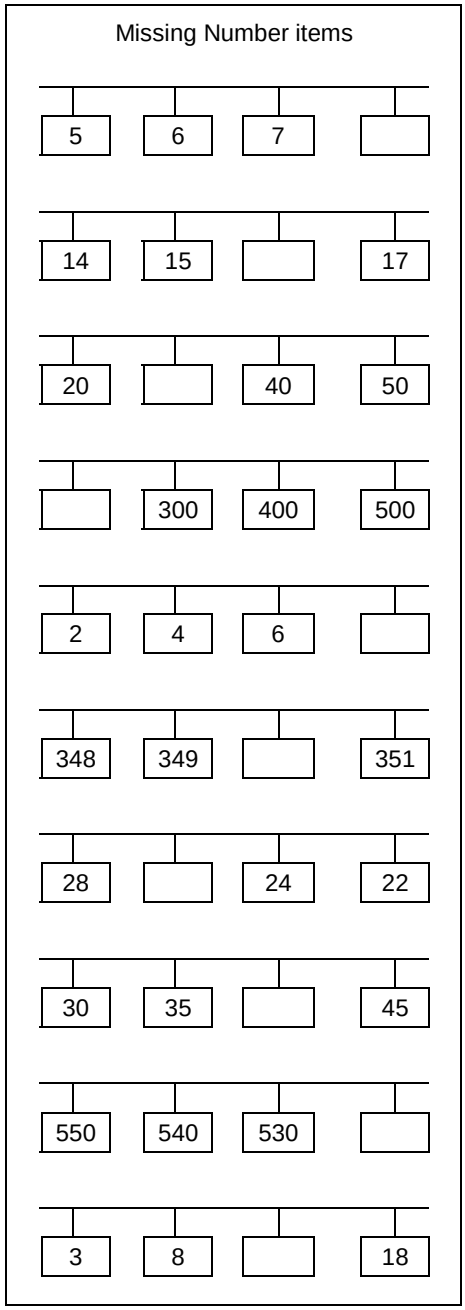
Although a rudimentary perception of magnitude is present early in life, further development in this subdomain is supported by education (Piazza et al., 2013). Non-symbolic and symbolic number-related thinking mutually support development in both. Number comparison has been shown to be a highly valid and reliable measure of early mathematical ability (Clarke & Shinn, 2004). The Number Discrimination subtest allows assessment of place-value skills. It was designed to test this ability through comparisons in which the higher number has lower ones and or tens digits than the lower number. Numbers that are further apart (e.g., 29 and 83) are generally easier to discriminate than those closer together (e.g., 32 and 29); therefore, numbers that are closer together have also been included to assess a broader range of abilities.

2.4.3 Missing Number

Description: The Missing Number subtest is an untimed test of 10 items with a stop rule after four successive errors. The items are presented as four horizontally aligned boxes, three of which contain numbers and one of which is empty (the target missing number). Eight of the items increase in number from left to right; two of the items decrease in number from left to right. Items 1, 2, and 6 increase by one (in a set of one-, two-, and three-digit numbers, respectively). Items 3, 4, 5, and 8 increase by tens, hundreds, twos, and fives, respectively. Items 7 and 9 decrease by twos and tens, respectively. The last item with numerals within the range of 1–20 increases by fives, but does not begin with a multiple of five. Students are asked to state the number that belongs in the empty box.

Background: The ability to detect number patterns (e.g., skip counting [also called count-bys]) of 20, 30, 40, and so on is an important early skill that can support later mathematical skills such as multiplication (Geary, 1994). However, the primary reason why mathematical patterns serve as foundational skills is their linkage to algebraic thinking (Sarama & Clements, 2009). As Sarama and Clements succinctly state, "…patterning is the search for mathematical regularities and structures, to bring order, cohesion, and predictability to seemingly unorganized situations and facilitate generalizations beyond the information directly available" (p. 319). Some researchers suggest that functional thinking should be taught in the early elementary years to support children's developing algebraic thinking (Blanton & Kaput, 2004). Examples of this are "T-charts" to visually represent patterns with, for instance, the number of dogs on the left, and the number of legs represented by those dogs (1 dog, 4 legs; 2 dogs, 8 legs; 3 dogs, 12 legs; and so on). Number patterns become evident through children's organization of the dog data: As the number of dogs increases by one, the number of legs increases by four.

Number of dogs	Number of legs
1	**4**
2	8
3	12
4	16



This functional thinking can be described as "…representational thinking that focuses on the relationship between two (or more) varying quantities" (Smith, 2008, p. 143). Skill in detecting number patterns also predicts later mathematical performance and shows improvement over the course of the school year (Clarke & Shinn, 2004), fulfilling two of the requirements for inclusion in the Core EGMA: predictive power and teachability.

2.4.4 Addition and Subtraction

Description: The Addition and Subtraction Level 1 subtests are timed tests (60 seconds) consisting of 20 items each that increase in difficulty. No addends are greater than 10, and no sums are greater than 19. The subtraction problems are the inverse of the addition problems. Three of the items mirror three of the Word Problems items. Assessors also keep track of whether the student used one of three problem-solving strategies: finger/tick marks, paper and pencil calculation, or mental arithmetic..

Sample addition and subtraction level 1 items

1 + 3 = ☐	4 – 1 = ☐
3 + 2 = ☐	5 – 2 = ☐
6 + 2 = ☐	8 – 2 = ☐
4 + 5 = ☐	9 – 5 = ☐
3 + 3 = ☐	6 – 3 = ☐

Addition and subtraction level 2 items

13 + 6 = ☐	19 – 6 = ☐
18 + 7 = ☐	25 – 7 = ☐
12 + 14 = ☐	26 – 14 = ☐
22 + 37 = ☐	59 – 37 = ☐
38 + 26 = ☐	64 – 26 = ☐

Description: The Addition and Subtraction Levels 2 subtests are untimed tests consisting of five items each that increase in difficulty, with a stop rule of four successive errors. Addition Level 2 is not given to students who receive a score of zero for Addition Level 1, and Subtraction Level 2 is not given to students who receive a score of zero for Subtraction Level 1. No sums are greater than 70. The subtraction problems are the inverse of the addition problems.

Background: Arithmetic (addition, subtraction, multiplication and division) serves as the foundation for the skills necessary in later mathematics and science education (Ashcraft, 1982). As noted in Section 2.1, children have rudimentary skills in arithmetic prior to enrolling in school. Studies show that children understand that adding to a small group of objects increases the quantity of items and that taking away from the set decreases the quantity (Starkey, 1992). In early arithmetic development, children use objects to represent addition and subtraction problems (Geary, 1994). With regard to problem-solving strategies, children frequently advance from a "count-all" strategy (counting all of the items from both sets) to counting on from the larger set (the "min" strategy). These strategies can be used for both object-based addition and symbol (numeral)-based addition.

In the Core EGMA, arithmetic skills are represented by the Addition and Subtraction Levels 1 and 2 subtests. The ability to add and subtract are thus important skills to measure, but to be optimally useful, students must be "fluent" in arithmetic facts. The word "fluent" is commonly described as both fast and accurate (Common Core Standards Writing Team, 2013). Fluency is a combination of "just knowing" (sometimes referred to as automatic retrieval), using number patterns, and using strategies (Baroody, 1987b; Common Core Standards Writing Team, 2013). Research supports the idea that fluency in the addition of one-digit numbers is a core skill for later mathematical development and predicts later mathematical performance (Baroody, 1987b; Jordan, Hanich, & Kaplan, 2003; Reikeras, 2006). The

extent of students' fluency and the ways in which they are fluent evolves through primary school. Eventually, many number facts can be automatically retrieved (although other strategies continue to be used for those facts not accessible through automatic retrieval), which reduces working memory demands in solving more complex problems (Tolar, Lederberg, & Fletcher, 2009).

The Core EGMA measures fluency through two means: accuracy (or correctness) and speed. Because more extensive mental calculations are needed to perform two-digit arithmetic, the Addition and Subtraction Levels 2 subtests are untimed. These levels assess students' more advanced arithmetic skills, involving both nongrouping and grouping problems.

2.4.5 Word Problems

Description: The Word Problems subtest is an untimed test consisting of six items each that increase in difficulty, with a stop rule of four successive errors. Three of these items use numbers that match three items from the Addition and Subtraction Level 1 subtest. Assessors also keep track of whether the student used one of three problem-solving strategies: finger/tick marks, paper and pencil calculation, or solved problem in his or her head. Students are also provided with counters that can be used to solve the problem.

The purpose for learning mathematics is to solve real-world problems, which are rarely, if ever, presented as stand-alone equations. Instead, they require interpretation of a problem and an understanding of the operations required to solve that problem. Word problems mimic, in a rudimentary way, these real-world situations. In classrooms where arithmetic is taught without conceptual understanding (e.g., limited to call-and-response instruction of addition facts), word problems can be difficult for children to understand and solve. In addition, children in classrooms where the language of instruction may not be the home language may experience difficulties interpreting word problems. As is noted in Chapter 3, children responding to the Word Problems subtest of the Core EGMA in the datasets used for the reliability and validity analyses had particular difficulty in solving even simple word problems with small quantities. More research needs to be conducted in this area to determine the causes of these difficulties.

Much research has been conducted on word problems (Carpenter, Fennema, Franke, Levi, & Empson, 1999). As one can imagine, different types of problems have differing difficulty levels. The six items in word problems represent the following problem types:

Problem Type	Example
Change: Result Unknown	Two children are on the bus, three more children get on. How many children are on the bus altogether?
Combine: Result Unknown	There are six children on the bus, two are boys. The rest are girls. How many girls are there on the bus?
Compare: Change Unknown	There are two children on John's bus and seven children on Mary's bus. How many children must join John's bus so that it has the same number of children as Mary's bus?
Change: Start Unknown	Five children get on the bus. Now there are 12 children on the bus. How many children were on the bus to begin with.
Sharing	Four children share twelve candies equally between themselves. How many candies does each child get?
Multiplicative	There are five seats on the bus. There are two children on each seat. How many children are on the bus altogether?

It should be noted that the word problems are presented to students orally to prevent the possible confound of literacy. If students were required to read the problem themselves, it would be difficult to determine whether an incorrect answer occurred because of a lack of mathematical or literacy skills. Counters and pens and paper are provided to students in case they would like to use them.

2.5 Mathematical Subdomains Not Covered by the Core EGMA

As stated above, one of the requirements of large-scale assessments is that they not be unduly long. Not only do long assessments cause fatigue for the students being assessed (and for the assessors) but they are also more costly. To avoid the negative factors associated with longer tests, the current version of the 20-minute Core EGMA does not cover a number of important subdomains that exist in many curricula for primary grades 1–3, specifically multiplication and division, fractions and decimals, geometry, measurement, and data analysis. Nor does it cover other subdomains that are predictive of later academic achievement, such as spatial relations and relational reasoning.

RTI is in the process of creating add-on subtests for those projects or countries that want to assess these skills.

2.6 Background of the Development of the Core EGMA

2.6.1 Item Development

Items in the EGMA[5] were developed by RTI in 2008 with the input of experts in the field and an in-depth literature review. The report titled *Early Grade Mathematics Assessment (EGMA): A Conceptual Framework Based on Mathematics Skills Development in Children* (RTI International, 2009a) reviews the literature on children's development of foundational skills related to number and operation. This literature review justified the development of the EGMA, including the mathematical subdomains it covers. The report also presented background on the cross-cultural evidence of commonalities in curricula. Curricula are generally based on an understanding of children's developmental progression within the specific domain. In the domain of mathematics, in particular during the early primary years, the components that constitute the foundation of later mathematics are well known (Clements & Sarama, 2007; Ginsburg, Klein, & Starkey, 1998; Nunes & Bryant, 1996). These components include the ability to name symbols used to represent numerosity (numerals) and discriminate between two quantities as represented by sets or numerals. These components also include the ability to detect number patterns, perform the operations of addition and subtraction, and apply these skills to real-life situations, generally assessed through word problems. This consistency in curricula supported the inclusion of the specific subtests in the EGMA. The original EGMA consisted of the following subtests:

- Counting Fluency (60 seconds)
- One-to-One Correspondence (fluency, 60 seconds)
- Number Identification (fluency, 60 seconds)
- Quantity Discrimination (untimed)
- Missing Number (untimed)
- Word Problems (untimed)
- Addition and Subtraction (untimed)
- Shape Recognition (untimed)
- Pattern Extension (untimed).

[5] Note that "EGMA" refers to the early versions of the instrument (2008 to September 2011) and "Core EGMA" refers to the version described in this Toolkit.

2.6.2 Expert Panel I

After the EGMA subtests were developed, a panel of university-based experts met in January 2009 to review and approve the EGMA for piloting. Members of this first panel, which we refer to in this toolkit as "Expert Panel I," were as follows:

- David Chard, Southern Methodist University
- Jeff Davis, American Institutes for Research
- Susan Empson, University of Texas at Austin
- Rochel Gelman, Rutgers University
- Linda Platas, University of California, Berkeley
- Robert Siegler, Carnegie Mellon University.

As can be noted in their biographies, the members of this panel possess a wide array of expertise in various areas, including international assessment, mathematical disabilities, curriculum development, and mathematical development in number and operation at population and individual levels. During the course of this meeting, Expert Panel I reviewed all aspects of the EGMA. This included the specific subdomains of mathematics covered, the scope and sequence of the items within the subtests, and discussion of any important subdomains not covered. Expert Panel I unanimously approved all of the subtests. Minor recommendations were made regarding the scope of the items in the Number Identification and Missing Number subtests. A majority of Expert Panel I recommended adding two new subtests: Number Line Estimation and Shape Attributes. The revised EGMA then consisted of the following subtests:

- Counting Fluency (untimed)
- Counting One-to-One Correspondence (untimed)
- Number Identification (fluency, 60 seconds)
- Quantity Discrimination (60 seconds)
- Number Line Estimation
- Missing Number
- Word Problems (untimed)
- Addition and Subtraction (untimed)
- Shape Recognition (untimed)
- Shape Attributes (untimed)
- Pattern and Number Extension (untimed).

2.6.3 EGMA Pilot in Malindi, Kenya

In July 2009, the EGMA, as approved by Expert Panel I, was piloted in 20 schools in Malindi District, Kenya (see Reubens & Kline, 2009). During the EGMA adaptation workshop in Kenya, the following changes were made to the panel-approved EGMA:

- Number Identification originally included numbers 1–99; however, the Kenyan curriculum in primary grades 1–3 covered up to 999. These numbers were added to the Number Identification subtest.
- Similarly, the addition and subtraction problems were deemed to be too easy; therefore, more difficult problems were added. A change was also made to time both the Level 1 and Level 2 Addition and Subtraction subtests to determine levels of fluency.

- The pilot assessments took much longer than anticipated (well over 20 minutes). To shorten the assessment time, Counting Fluency, One-to-One Correspondence, and Shape Attributes were eliminated from the pilot.
- Kenyan students had difficulty understanding the Number Line Estimation subtest. During the limited assessments carried out as part of the adaptation workshop, assessors spent a considerable amount of time trying to explain this subtest. Because of this difficulty, it was eliminated from the EGMA for the pilot.

2.6.4 Initial Implementation and Refinement of EGMA

The first formal implementation of the EGMA occurred in Liberia in November 2009. A five-day adaption workshop and training was held. This implementation was unusual in that it was to assess the mathematical skills of students in CESLY, an accelerated learning program for students ages 10-18, and ALPP Youth ages 15-35 who had been prevented from attending school during the 14-year civil war. Prior to implementation, a review of the curriculum and teacher manuals was conducted. Due to the mismatch of the EGMA in arithmetic to the curriculum for these two groups, two subtests (i.e., Multiplication and Division) were added. It is important to note that these students were not in grades typical for their ages. The curriculum was accelerated, but students were still at early primary grade-level reading and mathematics at the time of assessment. To shorten the EGMA to the allotted assessment time, the Word Problems subtest was eliminated. Again due to the mismatch, the Addition and Subtraction subtest were divided into Levels 1 (one-digit addends) and 2 (included two-digit addends).

This resulting EGMA consisted of the following subtests:

- Number Identification (20 items; untimed; stop rule for three successive incorrect answers)
- Quantity Discrimination (20 items; 60 seconds; stop rule if only the first three items were incorrect)
- Missing number (eight items; stop rule for three successive incorrect answers)
- Addition/Subtraction Level 1 (10 items each; one-digit addends and subtrahends [quantity subtracted]); 60 seconds; no stop rule)
- Addition/Subtraction Level 2 (nine items each; one- and two-digit addends and subtrahends; two minutes; stop rule for three successive incorrect answers)
- Multiplication/Division (six items each; division inverse of multiplication; untimed; stop rule for three successive incorrect answers)
- Shape Recognition (four shape sheets; untimed; stop rule for each sheet at the first incorrectly identified shape).

2.6.5 Field Use of the EGMA

Over the next several years, the EGMA was implemented in 14 countries. Although the EGMA covered a generally accepted scope and sequence of number and operation development, adaptations were essential at the local level to accommodate differences in language and dialect as well as cultural contexts. The EGMA was translated into those languages spoken in the classroom, and the written and oral instructions for the assessor and students were adapted to include idiom appropriate to the region and language(s) of instruction. The word problems were changed to reflect cultural context, so that they included objects familiar to the students (if students do not understand the vocabulary used, then comprehension becomes a confound in the measurement of mathematical skills).

Over time, other adaptations were made to the EGMA, including the addition of subtests covering other mathematical subdomains such as fractions, decimals, multiplication, and division. Number and

operations were expanded in both directions, with additions of subtests such as Counting and One-to-One Correspondence on the lower end and higher and more complex numbers in the Number Identification subtest and written addition and subtraction problems on the higher end. Although Ministries of Education benefited from the additional knowledge gained about students' mathematical skills because of these changes, it became more difficult to define the "standard" EGMA with so many versions in use.

Recall that the original purpose of the EGMA was a need for a reliable measure of early grade mathematical skills that could be used across countries. The instrument was designed to provide stakeholders, from Ministries of Education to aid agencies to local education officials, with the information essential to making informed changes in teacher education and support, curriculum development, and implementation at the local and/or national level. Because of this overarching purpose, the EGMA was designed to sample skills that might indicate the need for intervention. Additionally, it needed to be quick to administer, have high face validity, allow alternate forms for multiple administrations, and result in reliable and valid data. To this end, the following criteria had been established for subtests to be included:

- They must represent a progression of foundational skills that support proficiency in mathematics.
- Research must indicate that the subtests have predictive power—i.e., they must test skills related to students future performance in mathematics.
- They must be common in many curricula for early grades.
- They must be teachable.

Over the course of implementation from 2009 to 2011, due to the variation in subtests, there was concern that these original goals of the EGMA may have been compromised. Therefore, a decision was made to convene a second expert panel (henceforth referred to as Expert Panel II). Expert Panel II included many original members of Expert Panel I, as well as those who had conducted implementation efforts. Expert Panel II also included more experts with international experience. More details about Expert Panel II are presented in Section 2.6.5 below.

2.6.5 Expert Panel II

In light of this proliferation in use and application of the initial EGMA, RTI staff and consultants (including several members of Expert Panel I) decided to create the Core EGMA (based on the original EGMA subtests) with stable and research-based properties. Much data concerning children's performance on the assessment had been gathered during the years of implementation. To determine which subtests and items should be retained and/or included in such an assessment, Expert Panel II was convened. Expert Panel II consisted of several members of Expert Panel I (i.e., Drs. David Chard, Jeff Davis, Rochel Gelman, and Linda Platas), along with a highly skilled trainer from the field (Aarnout Brombacher), and a U.S. early mathematics expert (Dr. Douglas H. Clements). Expert Panel II also consisted of U.S. and international assessment experts (Drs. Albert Akyeampong, and Leanne Ketterlin Geller), and experts in cross-cultural mathematical development and teacher education (Drs. Jan de Lange, Terezinha Nunes, Annie Savard, and Alejandra Sorto). Of course, many of these individuals have expertise in more than one of these areas. Biographies of members of Expert Panel II can be found in Appendix A.

The meeting of Expert Panel II had two primary goals: to review and obtain agreement on a "core" instrument, including recommendations for refinement; and to obtain recommendations on potential "expansions" of the instrument and on addressing administration issues that had arisen. These goals were accomplished through whole-group discussions and by breakout teams consisting of experts in the subdomains covered by specific subtests. Integrated into these discussions were conversations on the

methodology of assessment, including the number of items, stop rules, and issues of timing for estimating fluency.

By the end of the meeting, the members of Expert Panel II had agreed upon eight subtests for the Core EGMA. These subtests were Number Identification, Number Discrimination (originally titled Quantity Discrimination), Missing Number, Word Problems, Addition Level 1, Addition Level 2, Subtraction Level 1, and Subtraction Level 2. In addition to the agreement on these subtests panel members also reached an agreement on the scope and sequence of the items in each subtest. The panel members documented the item specifications and assessment methodologies for each of the subtests, and then recommended the development of several additional "add-on" subtests: Relational Reasoning, Spatial Reasoning, and Money. This toolkit reflects the current state of the instrument.

2.6.6 Subsequent Implentation

The Core EGMA, as described in this toolkit, was implemented by RTI in the Dominican Republic and in Kenya under task orders of the EdData II project. Validity and reliability estimates were drawn from these implementations. More information about these estimates is discussed in Chapter 3.

Chapter 3: Technical Adequacy of the Core EGMA

"Validity" is defined as the trustworthiness and meaningfulness of the uses and interpretations of test results. Validity is not a property of a specific test; instead, this term refers to the appropriateness of the uses and interpretations of the data obtained from administering the test. As such, results from a specific test might be useful for making Decision A, but not for making Decision B. In this situation, it could be said that Decision A is a valid use for the test, but Decision B is not. To make this determination, the quality and plausibility of the evidence that supports a specific use or interpretation are examined. A valid use or interpretation is supported by converging evidence that forms a viable argument in favor of that use or interpretation. If the evidence does not converge, the test results may not be appropriate for the specified use or interpretation.

This chapter discusses the validity evidence for the uses and interpretations of the Core EGMA. A degree of technical knowledge is needed in order to interpret the findings and conclusions in this chapter.

To guide the collection and evaluation of the validity evidence, we used the description of validity evidence as presented in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). These standards specify five sources of evidence that might be useful for evaluating the validity of the test score uses and interpretations. These five sources of evidence include: tested content, response processes, internal structure, relations to other variables, and consequences of testing. The remaining sections of this chapter focus on presenting technical evidence based on the tested content (Section 3.1), internal structure (Section 3.2), and relations to other variables (Section 3.3). Currently, only initial evidence is available on response processes or consequences of testing. When these data are deemed final, RTI will share them with the international community and also propose future directions for research.

The results presented in this chapter are based on data sets from two countries. At the time of analysis, these were the only available data sets in which the core instrument was administered. These data sets had large amounts of zero scores. Therefore, due to the nature of these data sets, the results of the technical adequacy are limited. However, the RTI research team believes that there remains a need to make available the technical information, limited as it is. Over time, as more countries use the Core EGMA, further data sets will be available and further analyses can be conducted. It is the hope of this team that the international community will contribute to continual improvements in the EGMA instrument.

3.1 Evidence Based on Tested Content

Content-related evidence for validity examines *what* an instrument measures and *how well* it does so. Sections 2.4 and 2.6 provide support that the construct under investigation for each subtest of the Core EGMA is appropriate and representative of the content that is specified and sampled via the Core EGMA. As is noted in those sections, this evidence includes a systematic examination of the scope and sequence of the tested content in comparison with curricula and by subject matter experts.

3.2 Evidence Based on Internal Structure

Evidence based on the internal structure of a test is needed to verify that the tested content is consistent, and conforms to the construct, which forms the basis of test interpretation (AERA, APA, & NCME, 1999). The internal structure of the Core EGMA was examined at the subtest level. Specifically, the internal consistency reliability was examined at the subtest and item levels. In this section, findings are presented from classical test theory and item response theory analyses. Knowledge of psychometrics may be needed to fully interpret these data.

Data presented in this section were obtained from independent administrations of the Core EGMA in two countries. In Country 1, 4385 students participated in the assessment at the beginning of another project. Approximately half of the sample were in grade 1 (n = 2192); the remaining 2193 students were in grade 2. In Country 2, 1007 students participated in the assessment; 494 students were in grade 1. In grade 2, were 510 students, and 3 students were in grade 3. All data were collected in 2012 by trained assessors.

Subtest-level reliability. Cronbach's alpha was used to examine subtest-level reliability. Internal consistency reliability results with Cronbach's alpha estimates greater than .80 are suitable for making *summative* decisions. Because Cronbach's alpha is sensitive to missing data, only student records with complete data were included in the calculation for the Core EGMA subtests. Given the administration procedures for the Core EGMA subtests, including timing and stopping rules, many student records were excluded from the analyses. Based on the results presented in **Table 1,** the Core EGMA subtests with acceptable Cronbach's alpha coefficients for making summative decisions included Number Identification, Number Discrimination, and Addition Level 1.

Internal consistency reliability results with Cronbach's alpha estimates greater than .70 are suitable for making *formative* decisions. Core EGMA subtests with acceptable Cronbach's alpha coefficients for making formative decisions included the previously mentioned subtests and Subtraction Level 1.

Cronbach's alpha estimates less than .60 indicate that subtest results may not be sufficiently consistent for making educational decisions at all. The Core EGMA subtests with insufficient internal reliability evidence included Missing Number, Word Problems, Addition Level 2, and Subtraction Level 2. A possible explanation for these findings is the limited number of items included in these subtests. Estimates of Cronbach's alpha are sensitive to the number of items on the subtest: Subtests with a small number of items often have weak reliability estimate. The Word Problems, Addition Level 2, and Subtraction Level 2 subtests consist of 5 items each. As such, Cronbach's alpha may be improved by adding additional items. Other revisions to these subtests may be necessary to improve the internal consistency reliability.

Item-level reliability. Item-total correlations generated through Rasch analyses were used to examine item-level reliability. Item-total correlations examine the consistency between respondents' scores (ability estimates) and their answers to the item. Corrected correlations do not include the respondents' answers to the item in the estimate of their score. Rasch output also provides the expected item-total correlation as a reference point for interpreting extreme values. Item-total correlations should be greater than .20. All of the Core EGMA subtests had sufficient item-total correlations.

Rank order of item difficulty. Rank order of the item difficult was another dimension of the internal structure of the Core EGMA that was evaluated. The test specifications for all subtests state that the items should increase in difficulty from the beginning to the end of the test. To evaluate this specification, the item order was empirically compared to the item difficulty parameters estimated using

Rasch analyses. Spearman's rho correlation coefficient was calculated to determine the consistency in ranking between the theoretical item difficulty and the empirical estimates of difficulty. All of the Core EGMA subtests evaluated for this chapter, excluding Word Problems, had statistically significant correlation coefficients, indicating that the theoretical rank ordering was significantly related to the empirically observed rank ordering.

In summary, the Core EGMA subtests with sufficient convergent evidence indicating a strong internal structure included Number Identification, Number Discrimination, Addition Level 1, and Subtraction Level 1. Core EGMA subtests with moderate evidence supporting the internal structure included Missing Number, Addition Level 2, and Subtraction Level 2. The Word Problems subtest did not have sufficient evidence to justify the internal structure of the subtest. The usefulness of the Word Problems subtest for interpreting students' knowledge and skills in early grade mathematics should be examined given these results. **Table 1** presents the reliability estimates for each subtest.

Table 1. Reliability Estimates for Core EGMA Subtests

Subtest	Cronbach's Alpha[a]	Item-Total Correlation	Spearman's Rho
Number Identification	.94	.42–.77	.94[b]
Number Discrimation	.82	.58–.75	.95[b]
Missing Number	.58	.22–.79	.95[b]
Word Problems	.44	.35–.77	.70
Addition Level 1	.81	.58–.81	.76[b]
Addition Level 2	.68	.51–.83	1.00[b]
Subtraction Level 1	.79	.47–.85	.64[b]
Subtraction Level 2	.63	.56–.82	1.00[b]

[a] Values for Cronbach's alpha were estimated using Winsteps®. Values are estimates due to missing data.
[b] Data are statistically significant at the .01 level.

Recent Rasch analyses were conducted to examine item difficulty parameters. Results indicated that several items on each subtest have unacceptable fit statistics. There are several reasons that might explain these finding, including the limited range of mathematics ability in the tested sample and the limited number of items in each subtest. Additional analyses are needed to confirm these findings. Should the results be verified, revisions to the Core EGMA subtests may be warranted.

Total score possibilities. Another aspect of the internal structure of the Core EGMA to consider when evaluating the validity evidence for the Core EGMA is the level at which the scores are reported. Based on the original test specifications, results are reported for each Core EGMA subtest independently. The scores are reported as raw scores for all subtests except for the timed subtests (Number Identification, Addition Level 1, Subtraction Level 1), which are reported as a rate, meaning items answered correctly per minute. The subtest results are not interpreted in reference to a criterion or a normative range. Instead, the subtest results provide descriptive information.

In some instances, a total score might be requested. Because the scales on which each subtest is based are different, aggregating the subtest scores to report a total score should be performed with care. For the timed subtests, an average rate of correct responses per minute can be calculated. However, the utility of this rate for providing descriptive information should be considered. If a criterion for evaluating levels of proficiency is desired, then the predictive utility of the cut scores should be evaluated.

For the untimed subtests, the scores can be aggregated by determining the average of the proportion of items answered correctly. For example, the proportion of items correct for each of the untimed subtests (Number Discrimination, Missing Number, Word Problems, Addition Level 2, Subtraction Level 2) could be averaged to derive the average proportion correct. Again, the utility of this average proportion correct should be examined for providing descriptive information.

3.3 Evidence Based on Linkages to Other Variables

In general, validity evidence that evaluates the linkage between test results and other variables is used to substantiate claims about the underlying construct of the test. Evidence is gathered to examine the relationship of the test results to other tests that purport to measure the same or different constructs. To establish convergent evidence that the test under consideration is measuring a specific construct, the linkage between the test results and other tests that are designed to measure the same or similar constructs is evaluated. To evaluate the degree to which the test is not measuring a divergent construct, the linkage between the test results and tests of different constructs is considered.

The Core EGMA subtests are intended to measure different, but related, components of the construct of early numeracy and should be moderate to strongly related. Convergent validity evidence for the Core EGMA was examined by evaluating the correlation coefficients between the subtests within the two sample data sets described in Section 3.2 (see **Table 2**). Number Identification, Missing Number, and Number Discrimination had a moderate to strong correlation. Addition Level 1 was moderately correlated with these subtests, was moderate to strongly correlated with Subtraction Level 1, and was moderately correlated with Addition Level 2. Subtraction Level 1 was moderately correlated with Number Discrimination and Missing Number. Addition and Subtraction Level 2 were moderately correlated.

These data provide convergent evidence that all of the Core EGMA subtests except Word Problems, Addition Level 2, and Subtraction Level 2 measure related components of the construct of early numeracy. Addition and Subtraction Levels 2 were moderately related, which may indicate that they were assessing a unique component of early numeracy. Additional evidence is needed to evaluate the linkage between these subtests.

Table 2. Correlation Coefficients for the Core EGMA Subtests

	Number Identification	Quantity Comparison	Missing Number	Word Problems	Addition Level 1	Addition Level 2	Subtraction Level 1
Number Discrimination	.74[a]	—	—	—	—	—	—
Missing Number	.66[a]	.64[a]	—	—	—	—	—
Word Problems	.32[a]	.37[a]	.37[a]	—	—	—	—
Addition Level 1	.59[a]	.59[a]	.58[a]	.40[a]	—	—	—
Addition Level 2	.34[a]	.38[a]	.40[a]	.28[a]	.54[a]	—	—
Subtraction Level 1	.48[a]	.52[a]	.52[a]	.39[a]	.69[a]	.46[a]	—
Subtraction Level 2	.22[a]	.25[a]	.29[a]	.21[a]	.32[a]	.57[a]	.39[a]

[a] The correlation was statistically significant at the 0.01 level (two-tailed).
Note: Correlations were calculated using the examinees' raw scores.

Discriminant validity evidence is evaluated by comparing the test results with a test measuring a dissimilar construct. For the Core EGMA, the correlation between the subtests and a test of oral reading fluency was evaluated. Oral reading fluency was administered as a component of the Early Grade Reading Assessment (EGRA). Because mathematics and reading are theorized to be unrelated constructs, there should be minimal linkages between these tests. Based on the data presented in **Table 3**, a moderate to low linkage existed between oral reading fluency and Number Identification, Number Discrimination, Missing Number, and Addition and Subtraction Levels 1. Weak to low linkages existed between oral reading fluency and Addition and Subtraction Levels 2 and Word Problems. These findings provide evidence that the Core EGMA subtests are measuring a different construct than a test of oral reading fluency.

Table 3. Correlation Coefficients for the Core EGMA Subtests and Oral Reading Fluency

Core EGMA Subtest	Oral Reading Fluency[a]
Number Identification	.37
Number Discrimination	.36
Missing Number	.39
Word Problems	.20
Addition Level 1	.35
Addition Level 2	.23
Subtraction Level 1	.34
Subtraction Level 2	.14

[a] For all of the data in this column, the correlation was statistically significant at the 0.01 level (two-tailed).
Note: Correlations were calculated using the examinees' raw scores on the EGMA and EGRA.

3.4 Emerging Evidence Based on Response Processes and Consequences of Testing

As stated previously, sufficient data are not yet available to evaluate the validity evidence based on response processes or the consequences of testing. However, exploratory analyses are possible, given the available data from the administration of the Core EGMA in two countries (as described in Section 3.2).

3.4.1 Evidence based on response processes.

In general, examinees' response processes are examined to determine whether a test is assessing the construct to the level and depth in which it is intended. Most often, data from examinees are evaluated to determine whether the test items are eliciting the intended behaviors. Studies designed to gather these data include conducting interviews with examinees as they complete the test or interviewing examinees after they finish.

It is possible to explore the response processes in which students engage with the Core EGMA by examining the strategies students use when solving fluency-based items. The Core EGMA subtests that measure fluency include Number Identification, Addition Level 1, and Subtraction Level 1. Because measures that assess fluency in mathematics should be eliciting students' ability to quickly retrieve their mathematical knowledge, students should be using mental mathematics strategies. When RTI research teams examined the strategies students were using to solve the problems in Addition and Subtraction

Level 1, most students solved them in their heads (see **Table 4**). Some students used multiple strategies when adding and subtracting. These findings provide initial evidence that these Core EGMA subtests are eliciting the desired response processes for the assessed constructs.

Additional data can be gathered to verify the response processes that students use when engaging in these and other subtests of the Core EGMA. Possible methods include interviewing students after they solve fluency-based items, or collecting think-aloud data during the problem-solving process. The think-aloud method involves students orally articulating their cognitive processes as they solve problems so as to make their covert thinking process more explicit (Someren, Barnard, & Sandberg, 1994). The observer can then document and analyze the strategies and processes used to solve problems.

Table 4. Strategy Use for Solving Addition and Subtraction Level 1 Problems

Strategy	Addition Level 1 (%)	Subtraction Level 1 (%)
Solved in head	72.9	66.1
Solved with fingers	52.3	51.6
Solved with counters	1.8	2.4
Solved with tick marks	1.6	3.5

3.4.2 Evidence based on the consequences of testing.

Because validity is related to the trustworthiness and meaningfulness of the uses and interpretations of test scores, evidence about the consequences of these uses should be examined. Specifically, evidence is needed to justify that the test scores are adequate to warrant the consequences. RTI research teams approached the evaluation of consequential aspects of validity for the uses and interpretations of the Core EGMA by examining a possible unintended use. The Core is not intended to make cross-country comparisons; however, some researchers and/or policy makers may want to make such comparisons. Group-level differences between the two countries previously described in Section 3.2 were examined to determine whether the data are sufficiently adequate to support this unintended use.

RTI research teams used available data from two countries (Country 1 and Country 2) to compute inferential statistics. To evaluate whether group differences in student performance were statistically significant, a multivariate analysis of variance (MANOVA) test was used. MANOVA was selected due to the possible covariation between the subtests of the Core EGMA that should be accounted for when statistical significance is evaluated. As a point of caution, although *F* tests are typically robust to violations of the assumption of normality, the distribution of scores on some of the Core EGMA subtests were significantly skewed, which could have impacted interpretation of the MANOVA results.

Three effects were evaluated for each Core subtest: main effect for country (irrespective of grade), main effect for grade (irrespective of country), and the interaction effect of country by grade. Statistically significant multivariate main effects were observed for country (Wilks' Lambda = 0.50, F (8, 3147) = 390.35, $p < .01$) and grade (Wilks' Lambda = 0.79, F (8, 3147) = 103.81, $p < .01$). The interaction effect of country by grade was also statistically significant (Wilks' Lambda = 0.95, F (8, 3147) = 19.98, $p < .01$). Statistically significant univariate main effects were observed for country and grade for all Core EGMA subtests. When the RTI research teams controlled for Type I error, interaction effects between country and grade were statistically significant for Number Identification (F [1, 3154] = 69.36, $p < .01$), Addition Level 2 (F [1, 3154] = 7.61, $p < .01$), Subtraction Level 1 (F [1, 3154] = 23.45, $p < .01$), and Subtraction Level 2 (F [1, 3154] = 9.77, $p < .01$).

To better understand whether these differences were related to differences in students' knowledge and skills, or might have been caused by other factors, RTI research teams evaluated differential item functioning (DIF). DIF indicates that one group of examinees scored differently than another group of examinees when controlling for overall ability. For example, if examinees with the same ability estimates respond differently to a question based on their group membership (i.e., country, gender), then the item is functioning differently for the examinees based on their group.

For the Core EGMA, DIF was examined based on the countries (Country 1 and Country 2) where the test was administered. Because variability in students' item-level responses and overall performance is required to calculate DIF, RTI research teams did not examine Core EGMA subtests with floor effects (i.e., Addition Levels 1 and 2, Subtraction Levels 1 and 2, Word Problems). In addition, RTI research teams did not evaluate the timed tests (Number Identification, Addition and Subtraction Levels 1) because due to the timed nature of the subtest, the data sets were incomplete. As such, RTI research teams evaluated items in the Number Discrimination and Missing Number subtests for DIF.

Using the Mantel-Haenszel approach, a standard protocol for calculating DIF, RTI research teams evaluated items for the Number Discrimination and Missing Number subtests. The following criteria for determining the degree of DIF (Linacre, 2013) were used:

- **Moderate to large DIF:** A DIF value greater than or equal to 0.64 and a statistically significant chi-square statistic ($p < .05$)
- **Slight to moderate DIF:** A DIF value greater than or equal to 0.43 and a statistically significant chi-square statistic ($p < .05$)
- **Negligible DIF:** A DIF value less than 0.43 and a non-statistically significant chi-square statistic ($p > .05$)

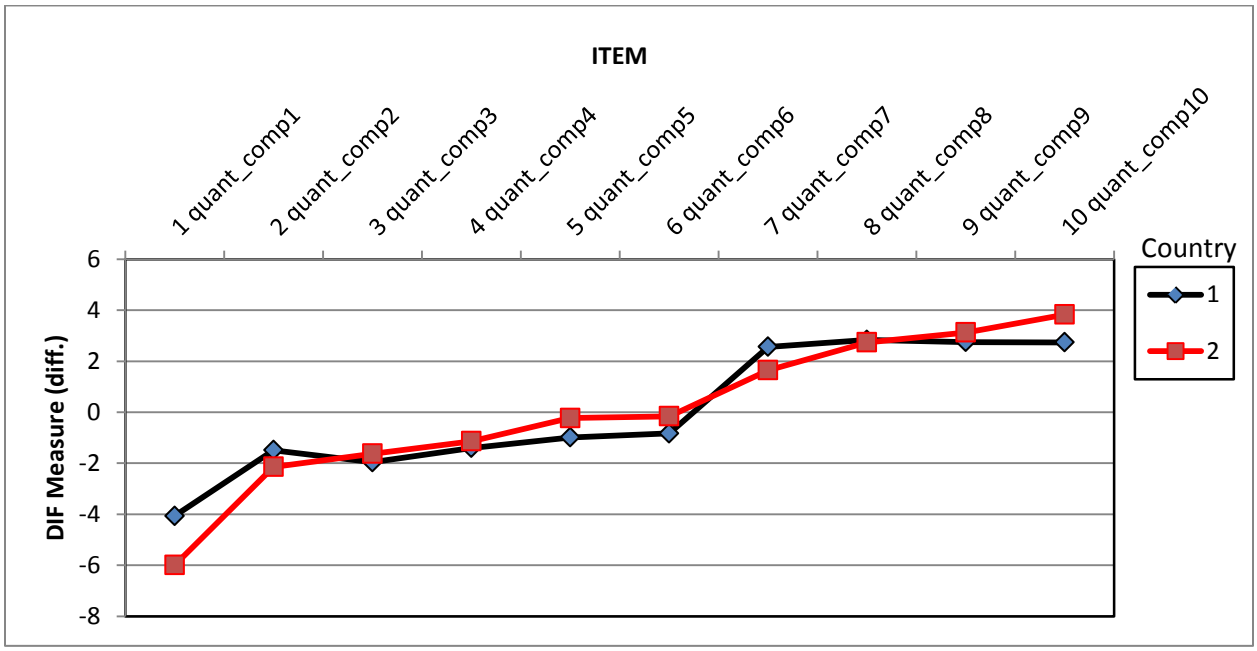
Results for the Number Discrimination and Missing Number subtests for Country 1 and Country 2 are presented in **Table 5**.

Table 5. DIF for Number Discrimination and Missing Number Subtests

Subtest	Number of Items with Moderate to Large DIF	Number of Items with Slight to Moderate DIF	Number of Items with Negligible DIF
Number Discrimination	5	3	2
Missing Number	1	1	8

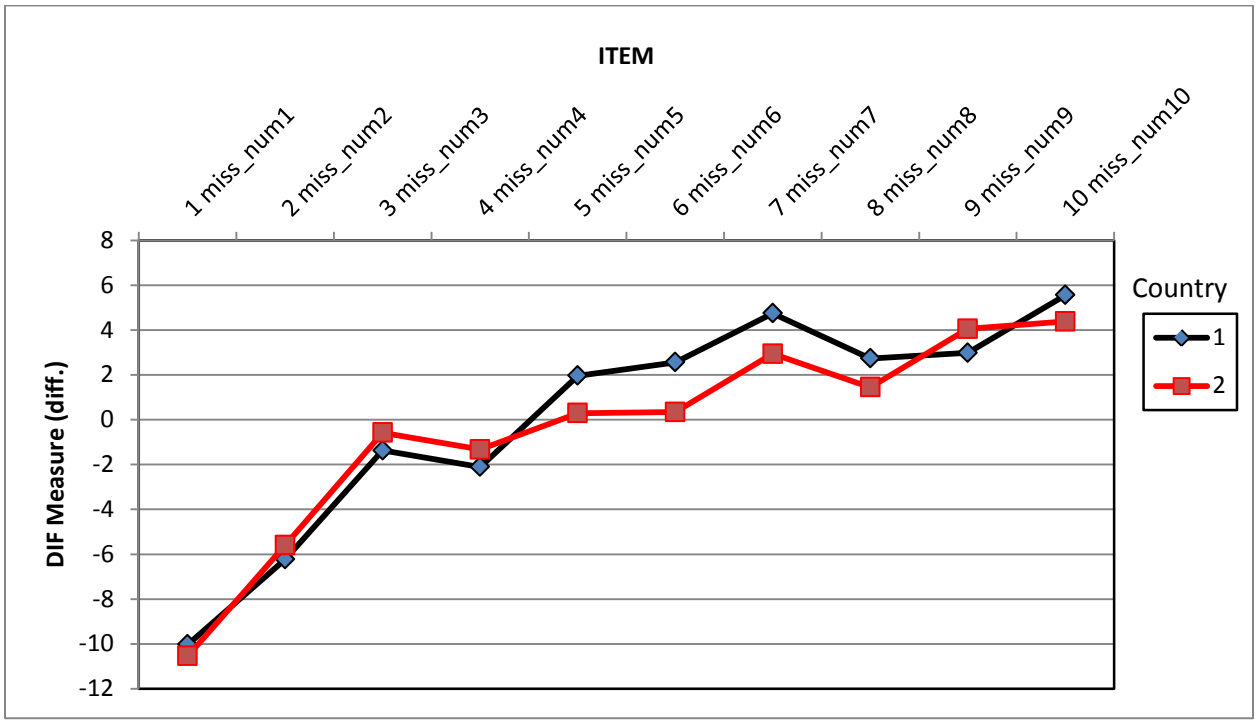
Graphical depictions of the DIF indicate the significance of the differential functioning by country. For Number Discrimination (see **Figure 1**), Items 1 (7 compared with 5), 2 (11 compared with 24), and 7 (146 compared with 153) met the criteria for Moderate to Large DIF in favor of Country 1. Items 6 (94 compared with 78) and 10 (867 compared with 965) met the criteria for Moderate to Large DIF in favor of Country 2. Items 4 (58 compared with 49), 5 (65 compared with 67), and 9 (623 compared with 632) met the criteria for Slight to Moderate DIF in favor of Country 2.

Figure 1. DIF for the Number Discrimination Subtest of the Core EGMA, by Country (1, 2)



For Missing Number, Item 9 (550, 540, 530, ___) was the only item that met the criteria for Moderate to Large DIF. For this item, the DIF was in favor of Country 2. Item 2 (14, 15, __, 17) met the criteria for Slight to Moderate DIF, also in favor of Country 2 (**Figure 2**).

Figure 2. DIF for the Missing Number Subtest of the Core EGMA, by Country (1, 2)



Data in Figure 1 indicate that for the Number Discrimination subtest of the Core EGMA, a significant number of items may function differently based on the country in which they are administered. Fewer items for the Missing Number subtest likely will function differently based on country. As such, there may be an element of the test or administration that impacts student performance based on the country of administration. These factors could include language of administration, country-specific curricular expectations, or other factors. These findings could have significant implications if data are used to compare performance across countries, because the test results may yield different information about students' knowledge and skills.

Note that when these Core EGMA subtests were examined for DIF by gender, no DIF was observed for the Number Discrimination and Missing Number subtests. As such, the items were not biased in favor of or against girls.

Overall, these results indicate that student performance on the Core EGMA (specifically the Number Discrimination and Missing Number subtests) should not be used to make cross-country comparisons. This finding confirms the stated purpose of the Core EGMA and provides evidence of potential negative consequences should the Core EGMA be used for this unintended purpose.

3.5 Conclusions Regarding Test Validity

In this section, RTI research teams discussed our examination of the validity evidence for the uses and interpretations of the Core EGMA. Specifically, the teams evaluated evidence based on content, response processes, internal structure, linkage to other variables, and consequences of testing. In general, the accumulated evidence pointed to the trustworthiness and meaningfulness of the results for making specific decisions to support student learning in early grade mathematics.

Chapter 4: EGMA Adaptation and Training

In this chapter, the RTI research team briefly outlines the adaptation and training processes for administering the EGMA. For additional information about these processes, please also see Chapters 4 and 5 of the EGRA Toolkit (RTI International, 2009b).

This chapter is presented as follows:

4.1 EGMA Adaptation Workshop
4.2 EGMA Assessor Training
4.3 Pilot study and instrument finalization
4.4 Other

This chapter assumes that the groundwork for the study has already been completed. That is the client and the service provider have agreed on the sample size, the variables that the data will be analyzed according to (e.g. male/female; urban/rural; private/public schools etc.), the timeframe for the study and the target audience of the results of the study.

4.1 EGMA Adaptation Workshop

An EGMA adaptation workshop typically follows five steps that are key to the successful development of the instrument. Guidance on these steps—intended for workshop design teams and facilitators—appears in Subsections 4.1.1 through 4.1.5.

4.1.1 Step 1: Invite participants

Participants in the EGMA adaptation workshop should include teachers, principals, district education officers, nonprofit workers in the field of education, experts in the development of early mathematics curricula, assessments, and education policies, and experts in the language that the EGMA will be administered in. To the extent possible, Ministry of Education officials should also be included, to build capacity to prepare, update, and implement such instruments. It is preferable if some of the assessors who will be gathering data also participate in the EGMA adaptation workshop.

Care should be taken in selecting the participants for the adaptation workshop. On the one hand, the participants contribute their local knowledge and expertise to the adaptation process and in so doing ensure that the EGMA is aligned with the national curriculum and is appropriate for the grades being assessed during the study. On the other hand, the participants, by their status within the education community of the country and within the mathematics community, validate the instrument.

4.1.2 Step 2: Clearly Define the Purpose of the EGMA

At the start of the adaptation workshop it is important for the leadership team and sponsors to define the purpose of the EGMA in this country including decisions that have already been taken in terms of sample characteristics (grades being assessed; urban/rural etc.). They should also understand their dual role, namely that of adaptation as well as that of validating the instrument for the country.

4.1.3 Step 3: Provide Background Information about the EGMA

The EGMA adaptation workshop facilitator should give all the attendees an overview of the entire instrument, including detailed information about each subtest of the EGMA. This information includes the theoretical and research basis for each subtest and item, which can be found in Chapter 2 of this EGMA toolkit. Facilitators of the EGMA adaptation workshop should also discuss the rationale for the item specifications and basic information about scoring. It is important to remember to clearly communicate the purpose of the EGMA, as already determined, to the participants.

The EGMA assessment is typically unfamiliar to the participants in terms of the oral format of the assessment. Participants may at first express concerns about the assessment being "too easy" and/or about how children will/will not respond. For this reason, there is much to be gained by arranging a school visit during which the administration of the EGMA is demonstrated to the adaptation workshop participants. It works well to have one of the facilitators working with a teacher administer the assessment to a few children one after the other while the adaptation workshop participants sit behind the child being assessed. Irrespective of the school visit, it may also help to play a few videos of administrations in other countries and also to show the kinds of results that have typically emerged in other countries (without necessarily mentioning the countries) during the background session.

4.1.4 Step 4: Adapt the Instrument

Before starting the adaptation of the instrument, it is important to clarify the expectations of the participants in terms of what it means to adapt the EGMA instrument. These include but are not limited to:

- The EGMA consists of a range of subtests which taken together give a picture of the state of early grade mathematics learning. These subtests have been developed with care. **The role of the adaptation workshop is not to create new subtests and/or to redesign the existing subtests.** That said, the role of the adaptation workshop is to agree on the range of subtests (from the Core EGMA and the additional tasks being developed) to be piloted and the basis on which the final selection of subtests will be made. For example, in some countries the number identification subtest has been dropped from the assessment because, during the pilot, the subtest had a strong ceiling effect (many students answered all items correctly) – that is, there was not much to be learnt from administering the subtest. In a different context a pilot may reveal that performance on the number identification subtest is so weak that it is a good idea to add a counting subtest in the EGMA study.
- There must be alignment between the items in the EGMA and the country's curriculum for the grades being assessed. However, this does not mean that all of the topics in that country's curriculum will be assessed – see the design rationale for EGMA discussed in Chapter 2. Similarly, there may be items in the EGMA that do not appear in the national curriculum of the country; this does not mean that the items will be removed from the EGMA. The reason for keeping such items would be because research has shown that these are foundational skills predictive of future success in mathematics and assessing these skills may suggest the need to update the national curriculum.
- The language used in the instructions must be appropriate to the context. In particular, the contexts for the word problems must be appropriate and meaningful to the children being assessed.
- The EGMA expert panel has developed item specifications for the items in each of the subtests. These have been developed to ensure a trajectory of increasing conceptual demand across the items in a subtest as well as a broad assessment of the skills being assessed by the

subtest. For this reason it is important that the participants in the adaptation workshop first consider the rationale for each of the items in each of the subtests before suggesting changes.

It is important that the participants begin the adaptation of the instrument against this background. Once the ground rules for the adaptation process have been established the participants should begin the review of the EGMA subtests and items on a subtests and item by item basis. This will typically take no longer than 2 days in the case of EGMA.

4.1.5 Step 5: Create the Instrument in Electronic Format Using the Tangerine® Platform, or in Paper Format

EGMA data can be gathered through paper instruments or on mobile devices such as tablets, using the RTI-developed Tangerine® software (see www.tangerinecentral.org).[6]

Irrespective of how the EGMA data will be collected (by means of paper or using a data entry interface such as Tangerine® on a tablet) a paper copy of the instrument will have to be finalized so that if the data collection team experiences difficulties with the tablets then they can administer the assessments using paper and pencil. Templates for the EGMA paper instrument and stimulus sheets exist and should be used at this point in the adaptation workshop.

4.2 EGMA Assessor Training

EGMA assessor training seeks to ensure that assessors are comfortable using the instrument, that they possess the skills needed to make students feel comfortable with taking the EGMA, and that the assessors can successfully track large amounts of data. Three steps involved with EGMA assessor training and a pilot study are discussed in Subsections 4.2.1 through 4.2.3.

4.2.1 Step 1: Invite Participants

As with the adaptation workshop, participants in the EGMA assessor training should include teachers, principals, district education officers, and nonprofit workers in the field of education. To the extent possible, Ministry of Education officials also should be included, to build their capacity to continue this type of assessment on their own in the future.

The number of assessors to be trained is determined by the study design. The number of schools to be visited, the number of children to be assessed in every school and the time frame for the study all impact the number of assessors to be trained. For example, say that 160 schools are to be visited and 20 children to be assessed at every school (a total of 3,200 children) and that the study is to be completed in 10 days. There will to be at least 16 assessor teams (160 ÷ 10 = 16) – each team visiting one school per day (in some contexts there may be a need for more teams since the distances that some teams must travel will prevent them visiting a school each day). Assuming that each assessment takes 25 minutes (20 minutes of assessment and 5 minutes of swapping over from one child to the next), and assuming further that the school day is realistically 4 hours long, then each assessor can interview between 8 and 9 children on a day. To be safe each assessor team should consist of 3 assessors (3 × 8 = 24). So, for this example a minimum of 48 (16 teams × 3 assessors per team = 48 assessors) need to be trained. However, because there will be some attrition due to illness and other personal affairs it is probably wise to over train by between 10% and 20%, so in total 53 to 58 assessors should be trained. Furthermore, it is important to over train the number of assessors because, during the training, all assessors will be tested on their ability

[6] Tangerine® is open-source data collection software developed by RTI and customized for the EGRA and EGMA.

to conduct the assessment (see discussion of IRR in section 4.2.3) and some assessors may not be retained after the training. *Note*: the assumption that the assessment will take 25 minutes per child is realistic only if EGMA is the only assessment being conducted, if a pupil questionnaire is added and/or an EGRA is also being conducted then administration time and the number of assessors will increase.

When contracting the assessors it is very important that they are aware of the following points:

- **Time frame** - every assessor must be available for *all* of the following assessment activities:
  - The training days – typically about 5 if the study involves only EGMA, up to 8 days if there are other components.
  - The pilot study day(s) – depending on how the pilot study is planned this may involves some or all of the assessors for up to 2 days.
  - The study days – in the example above 10 days.
- Expectations with respect to **travel**.
- Finally, it is important that the assessors to be trained are informed from the start that they will be assessed on their **ability to conduct the assessment**, both by means of IRR and direct observation. Those assessors who do not achieve a minimum of 90% agreement with the standard during IRR and/or who consistently do not conduct the assessment properly when observed using the observation checklist (see Appendix A) will not be retained for the study.

It is also important to recruit assessors who can communicate effectively with young children and who have excellent organizational skills:

- **Rapport with children –** Assessors will be assessing young children individually. Therefore, they must be able to act in a child-friendly manner and put children at ease during the EGMA. Children who are fearful of assessors may not perform well.
- **Organizational skills –** Whether using paper or electronic data collection, assessors will need to keep track of large amounts of data and be able to follow very specific instructions on how to properly administer the assessment.

4.2.2 Step 2: Conduct the training

The major goals of the assessor training workshop are as follows:

- **Review underlying principles of the EGMA and the purpose of this particular EGMA administration**
- **Learn the administration rules for each of the subtests in the EGMA – see notes below** on typical rules for each of the subtests for the Core EGMA
- Provide assessors with extended opportunities to **practice administering the assessment**, including at least once, but preferably more often at a local school.
- **Train team supervisors on how to conduct the study at a school** – the schools visit(s) during the training workshop provide a good opportunity for team supervisors to practice these skills. Note, the general nature of the study (i.e. EGMA only or EGRA and EGMA and pupil questionnaires etc.) will determine how much work the team supervisor will have and how many children they can assess. Typically, if the study is only an EGMA study, the team supervisors can conduct as many assessments as each of the assessors. This manual will need to be adjusted for the context of the study but provides enough detail with respect to the issues to be dealt with by the supervisor.
- **Conduct a pilot study of the instrument** – This may or may not happen during the assessor training week itself, in some cases it will happen a few days or even a week or more later.

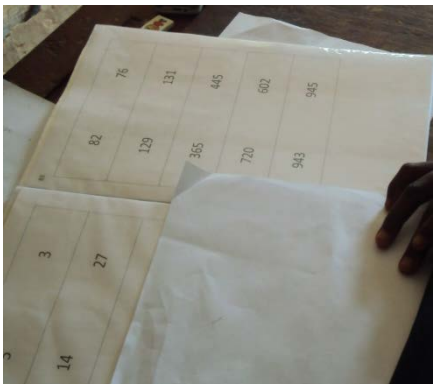
- **Finalize the instrument based on the results of the pilot study** – This is not strictly part of the assessor training workshop, but the project manager needs to build in time after the pilot study and before the main study to finalize the EGMA instrument based on the pilot study data.

EGMA test administration guidelines

As much as the administration rules are implied by the detail provided in the paper version of the instrument, these will need amplification and discussion during the training of the assessors. There are also a number of questions that will arise during the training; the comments below should address many if not all of these.

General comments

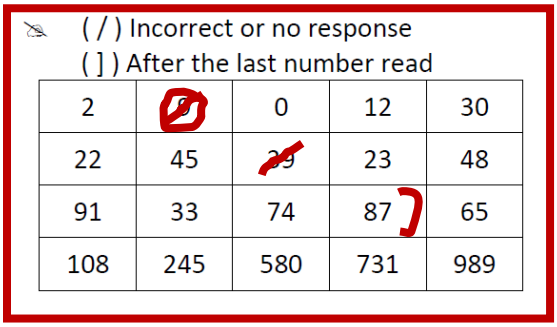
- **Children pointing at items** – It is important that the children are reminded by the assessor to point at the items as they respond.
  Rationale: As the assessor coordinates the child and the tablet (or paper version) on which they are recording the responses, it is possible for the assessor not to notice if a child goes back to an earlier item and auto-corrects. By making the child point to the item that they are responding to we reduce the possibility that the assessor scores the wrong item based on the response of the child.
- **Using a piece of paper to manage the child's progress on untimed subtests** – Assessors should anticipate that some children will benefit from using a piece of paper to cover the items on the stimulus sheet and exposing them one by one (this does not apply to any timed subtests). Typically, the child is able to move the piece of paper themselves as they move through the items.
  Rationale: For some subtests, in particular the *Number Discrimination* and *Missing Number* subtests, some children get overwhelmed by the successive items and exposing them one at a time can help.



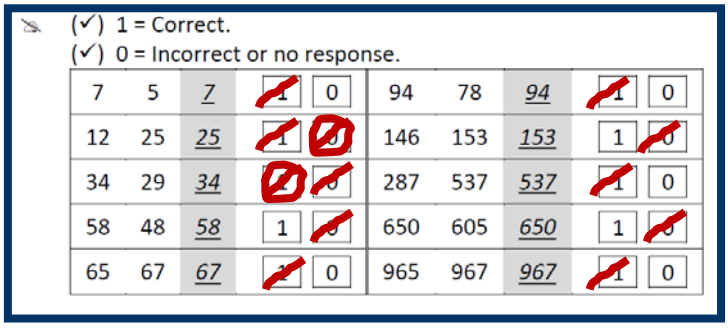
  Managing the nudge rule – For each subtest there is a nudge rule which in most cases involves encouraging the child to move to the next item if they have not responded to an item after 5 seconds.
  Rationale: The purpose of the nudge rule is to ensure that a child does not get stuck on one item and not attempt any others – this is especially important in the timed subtests (of which there are only three in EGMA). The nudge rule is also there to ensure that children do not take too long on each item and in so doing increase the test administration time beyond the target 15 to 20 minutes – this is especially true for the untimed tasks. In EGMA (unlike EGRA), however, the items increase in difficulty throughout each subtest and so nudging the child too quickly will only force them to attempt more difficult items where they are already taking their time (or struggling) on easier items.
- **Responding to the child** – Assessors need to be careful when they respond to children's answers. Assessors must never say "correct" or "incorrect". In fact, it is better that assessors say nothing as their responses will delay the progress of the child and increase the time taken to complete the tasks and assessment in general – this is a particular problem on the timed tasks as it may impact on the child's fluency score.
  Rationale: The reason for not saying "correct" and "incorrect" is very simply that when the child hears "incorrect" they will firstly want to try the item again until they get it right and

secondly hearing "incorrect" repeatedly is not good for the child's sense of self and may discourage the child in general.

- **Sticking to the script and use of local dialects** – Both the paper and Tangerine® instruments detail very carefully the words that the assessor must use when giving the instruction for each subtest. This should be regarded as a script – the assessor must say these words exactly as they are written in the instrument. While it is expected that the assessors will learn these words (through frequent use), in the same way that an actor learns the words of a script, and will not need to read them every time they administer the assessment, the assessors must continue to use the words as they appear in the instrument. **Use of local dialects** – in some contexts it may happen that despite all the care taken during translation to translate the instructions as precisely as possible, the assessor will find that the local dialect spoken by the children uses a few different words. In these cases the assessor may substitute the more familiar words in the local dialect to ensure that the child understands the instruction. In so doing, the assessors may not, however, change the meaning of the script and may certainly not embellish the script by, for example, adding hints etc. These changes should be conveyed to the team supervisor, which is useful in understanding any score differences that might arise though differing scripts.
  Rationale: It is important that every child in the study experiences the assessment in the same way for this reason we want assessors to "stick to the script". This cannot be stressed enough during training. Both during training and in the field, assessors need constant monitoring in this regard.
- **Self-correction by children** – A child may self-correct their answers. If a child gives a response to an item and then notices that he/she is not happy with that answer and then changes his/her answer, the last answer given by the child will be treated as the child's answer (even if the first answer was correct and the second answer incorrect). In the case of the paper instrument corrections by the child are recorded by drawing a circle around the first response/mark by the assessor. In the illustration of the Number Identification subtest (timed), the child first gave an incorrect response for item two and then self-corrected, so the circled stripe indicates that the response of the child is correct. In the illustration of the Number Discrimination subtest (untimed) the child self-corrected the answer for the second item and hence the assessor has circled the stripe indicating that the child gave an incorrect answer and has also indicated that the child answered correctly. In the case of the third item, the child first gave a correct answer and then changed her answer the final answer being incorrect. NOTE: if the assessor has not yet marked the instrument at the point that the child self-corrects there is no need to mark the item to indicate the self-correction, i.e. the assessor can simply record as if the child only gave one answer. In the case of Tangerine® the assessor simply changes the response on the tablet.


(/) Incorrect or no response
( ]) After the last number read

2	0	0	12	30
22	45	29	23	48
91	33	74	87]	65
108	245	580	731	989


(✓) 1 = Correct.
(✓) 0 = Incorrect or no response.

7	5	7	1	0	94	78	94	1	0
12	25	25	1	0	146	153	153	1	0
34	29	34	0	0	287	537	537	1	0
58	48	58	1	0	650	605	650	1	0
65	67	67	0	0	965	967	967	1	0

Rationale: We allow children to self-correct because we want to capture the best possible answer the child has.
NOTE: Children do not need to be told that they are allowed to change their responses (to self-correct). Children will spontaneously do so.

- **General arrangement of assessor and child** – In setting up the assessment space, there are two common arrangements. One is it the assessor and child sitting opposite each other. That way the assessor is able to hold the clipboard (in the case of the paper instrument) or tablet in such a way that the child cannot see the instrument that the assessor is working with. The other is the assessor and child sitting diagonally to each other (perhaps on the end of a table or bench). This allows the assessor tomore easily watch the child as he/she uses a finger to point to the items on the subtests. Regardless of the seating arrangement, it is important that the stimulus book is oriented in such a way that it is the correct way up for the child.
  Rationale: Children are curious and watch the assessor very carefully. We find that when the child can see what the assessor is doing (on the clipboard or tablet) they may wait with responding to the next item until the assessor has entered a response for the current item. In the case of the timed tasks the assessor only records incorrect responses and so this creates all sorts of problem. Furthermore some children may try to work out what the correct answer is by looking at the assessor's instrument.
- **Response language of the child** – In the case of multilingual contexts the child may respond to the items in any language that he/she chooses (provided of course that the assessor understands the response). In most applications in multilingual contexts the EGMA assessment will ask the assessor to record all the languages used by the child in responding to the each subtest. This information is analyzed to see if there are language use patterns associated with stronger and weaker performance
  Rationale: The EGMA assesses the child's knowledge of foundational mathematical skills, not their use of language. Furthermore in many multilingual contexts we find that while the language of instruction may be one language (say Kiswahili), the teacher actually teaches mathematics in English. In such cases we would rather that the child responds in English – the language in which he/she knows mathematics, rather than in Kiswahili. Remember, the purpose of the EGMA is to assess children's early mathematics competencies, not their language competencies.

Number identification subtest

- **Three-digit numbers** – in the case of the three digit numbers (and as far as it is appropriate to the language of the assessment) the child must be heard to say the word "hundred" when saying the number name. For example in the case of 731:
  - Seven hundred and thirty-one – correct
  - Seven hundred thirty-one – correct
  - Seven thirty-one – incorrect
  - Seven three one – incorrect

  Rationale: We are assessing foundational skills in mathematics and while "seven thirty-one" might be used colloquially, it does not convey the understanding of the meaning of the symbol 731 that is required for the child to be able to perform arithmetic with three-digit

numbers. By recording "seven thirty-one" as incorrect we may, during the item analysis, pick up a pattern of errors on three digit numbers and provide guidance in this regard.

Number Discrimination subtest

- **Saying the larger number** – The child must say the number that is larger. It is not enough for the child to point to the larger number and say "this one".
  Rationale: In the design of the subtest there are 10 items, in five cases the left hand number in the pair is larger (correct) and in five cases the right hand number in the pair is larger (correct). If the child did not engage properly with the task but simply pointed to one of the two numbers and said "this one" there is a good chance that he/she may get at least 5 correct responses and yet not know what he/she is doing. By asking the child to say the larger number we hope to reduce this possibility.
- **Three-digit numbers** – In the case of the three digit numbers in the quantity comparison subtest we do not mark the response as incorrect if we do not hear the word "hundred". For example in the case of 146 and 153, if the child responded "one five three" we would mark the response as correct.
  Rationale: In the number identification subtest "one five three" is marked as incorrect for the reasons discussed earlier. In the quantity comparison subtest we do not penalize the child again – in this subtest we are not assessing number identification.

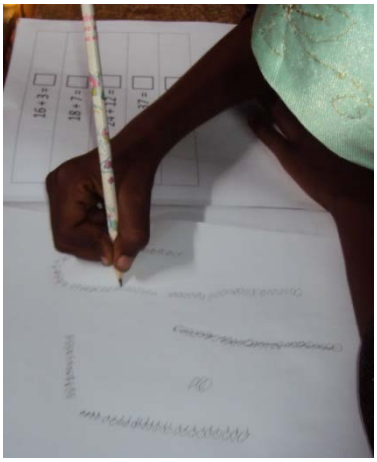
Missing Number subtest

- **Saying all the numbers in the pattern** – Some children will say all of the numbers in the pattern, for example "5, 6, 7, 8 – 8" to indicate that the missing number is 8. Children do not have to say all the numbers in the pattern; it is enough for them to say the missing number. Some children will say all the numbers because in the practice item (if they made a mistake) the assessors will have said "The number 3 goes here. Say the numbers with me. [Point to each number] 1, 2, 3, 4. 3 goes here." And so they think that they have to say the numbers. If a child says all the numbers in the pattern before saying the missing number, the assessor can gently say to the child once "You do not have to say all the numbers, you can simply give the answer if you like". If the child persists in saying all the numbers then the assessor should not stop the child from doing so.
  Rationale: Saying all the numbers increases the response time of the child. While this does not matter as much in the untimed tasks as it does in the timed tasks, it nonetheless increases the test administration time. If after being told that they do not have to say all the numbers the child continues to do so, we allow them to do so because it may be that the child needs to hear the pattern to work out the missing number.

Addition and subtraction L1 subtests

- **Repeating the problem** – Some children will repeat the addition/subtraction problem as part of their answer, for example "five plus three equals eight, eight". It is enough for children to give the answers only. If a child repeats the addition/subtraction problem as part of giving the answer, the assessor can gently say to the child once "You do not have to repeat the question, you can simply give the answer". If the child persists in repeating the question as part of their answer then the assessor should not stop the child from doing so.
  Rationale: repeating the addition/subtraction problem in this timed task uses up valuable time and reduces the child's fluency score, for this reason we would prefer the child to give only the answers. However, some children will persist in repeating the question and if they do so the assessor must not stop them from doing so – it may be that the child needs to hear the problem to help them recall/remember the answer. This is especially true of the items in the L1 addition and subtraction subtests.

Addition and subtraction L2 subtests



- **Using an inefficient strategy** – According to the nudge rules for these subtests, we nudge the child:
  - If the child uses an inefficient strategy (e.g. tick marks), ask the child "Do you know another way to solve the problem?"
  - If a child continues to use an inefficient strategy or stops on an item for 5 SECONDS.

  These rules should be treated as follows:

  - If the child does nothing in response to the item, the assessor will ask the child to move to the next item after 5 seconds.
  - If a child uses an inefficient strategy (e.g. tick marks), the assessor asks the child if they know another way to solve the problem. If the child says yes, the assessor should encourage them to use that method. If they say no and/or continue to use the inefficient strategy (e.g. tick marks) then the assessor will ask the child to move to the next item after 5 seconds.

  Rationale: Many children will be able to answer the addition and subtraction L2 items using an inefficient strategy (the most common of these is using tick marks). However, it takes a great deal of time and will often involve errors due to inaccuracy and it does not show an application of the foundational skills (assessed in the L1 subtests) to more complex problems. For this reason we will nudge the child to the next item after 5 seconds. If the child can determine an answer using the inefficient strategy in 5 seconds then we will accept the answer.
- **Application of the 5 second nudge rule** – The addition and subtraction L2 subtests provide a very good example of how the "5 seconds" in the nudge rule should be used. If the child is using an efficient strategy and appears to be working out the answer, the child may need more than 5 seconds to complete the problem, and should be allowed to do so. If the child is productively busy on a problem after a long time (the assessors will develop a feel for this) then the child can be asked to move to the next item.

  Rationale: If the child is productively busy with these problems, allow them more time to complete them; some children will need more than 5 seconds to calculate their responses to the items and we are more interested in their responses thane we are in the fluency with which they produce them. Of course, we must also balance completing the EGMA assessment in 15 to 20 minutes with allowing each child as much time as they want to calculate each answer.

Word Problem subtest

Presenting the problems to the child – Great care must be taken in presenting the problems to the child. In the general instruction to the child the assessor says: "Listen very carefully to each problem. If you need, I will repeat the problem for you." Ideally, we do not want to have to repeat the problem since this will take up a lot of time. For this reason, the problems have been set out as a series of short phrases each followed by the instruction [pause and check]. The expectation is that the assessor will use a conversational tone to present the problem one phrase at a time and then pause to look at the child waiting for some evidence that the child has understood each phrase. This does not mean that the assessor must ask: "Do you understand" although he/she may choose to do so. The assessor should definitely not ask

"How many children were on the bus?" or other such question following each phrase; that is, they should not ask a comprehension-like question about each phrase.
Rationale: Unless the child understands the story/situation presented in the problem they cannot make a plan to solve it. For this reason, the assessor must take as much care as possible to present the problem in an engaging manner. That said, if reasonable care has been taken in presenting the problems (as described above) and the child does not appear to be able to get started with the problem then the assessor must mark the child's response as incorrect and move to the next item.

- **Application of the nudge rule** – For the word problem subtest the nudge rule is a little different and should be applied as follows:
  - If the child does nothing in response to the problem then the assessors says "Let's try another problem" after 5 seconds (and marks the child's response as incorrect) .
  - If the child responds to the question by say: using their fingers to model to situation, or using the counters to model the situation or using the paper and pencil, and if the child continues to be engaged with the problem then the child can be allowed to continues in this way for up to one minute before the assessors says "Let's try another problem" (and marks the child's response as incorrect).

  Rationale: Solving the word problems will typically take the child a little longer. This is because the child must first make sense of the situation (problem), must then make a plan, must execute the plan and finally produce an answer/response. For this reason, we allow the child up to one minute to solve each problem, provided that they are visibly and productively engaged with the problem. Once more if the child produces a correct answer after one minute and two seconds then the answer will be accepted – the one minute rule (as with the 5 second rule in the other subtests) is a guide for what is considered a reasonable amount of time.
  NOTE: very few children actually need as much as one minute to provide an answer.

4.2.3 Step 3: Test for Inter-rater Reliability

Inter-rater reliability (IRR) is a method used as part of a study design to determine how accurately assessors are coding information when compared to an agreed standard. Assessor performance is measured against this standard. Using IRR test information clarifies how well the assessors are learning to code data during a data collection exercise and informs study organizers of overarching areas of concern so they are able to address any gaps in training. Additionally, the IRR provides the criteria to determine who among a pool of assessors should be selected as assessors for a particular study, ensuring that the most reliable assessors are selected to collect data for the study.

Typically, IRR tests take place three times during the training workshop to track how well the assessors are learning the coding requirements of the various subtests, while providing the opportunity for assessors to demonstrate improvements to existing skills refined during the workshop and the school visit practice sessions. It is important that each assessor receives individual feedback on their performance after each IRR test – this will enable them to work on areas of weakness.

In addition to highlighting areas that assessors should focus on as they practice, general error patterns across the assessors identified during the IRR tests suggest areas that may require specific or additional training during the assessor training week.

The IRR tests typically take place en masse to test all assessors on the same student responses. Before the start of the IRR test, the technical expert/facilitator (the "assessor") and a local expert (the "student") meet to design the responses by the student to the EGMA during the IRR test. The design of the responses helps to ensure that the IRR test will highlight typical errors the students will make during an assessment, the responses will also assess the assessors attention to important administration rules (discussed above). For example the responses during the IRR may include the "student" responding "seven three one" to the item 731 in the number identification subtest where it must be marked as incorrect. This set of responses constitutes the Gold Standard for each IRR test and serves as the standard against which the assessors who take the IRR test are scored.

The format of each IRR test typically follows the same pattern. A mock assessment is conducted at the front of the training room while the assessors observe and code the data from the staged assessment in much the same way as they would during a school visit for data collection (this will happen either on paper or using the tablets and Tangerine® software). The mock assessment involves the technical expert/facilitator as the "assessor" and the local experts as the "student." The mock assessment should be amplified so that all the assessors can hear the responses. If possible, it also helps to project the student stimulus booklet using a video camera and data projector, and for the "student" to point at the items as a student in the field would. In this manner, the assessors can all simultaneously follow along with the audio and visual process of the assessment just as they would during data collection. In this way, the IRR test evaluates the assessors' ability to use the paper/tablet appropriately, monitor students' responses, code students' responses accurately according to the design of each individual subtest protocol, and manage the audio, visual, and mechanical aspects of the assessments.

On the completion of the IRR test the responses of the assessors need to be uploaded into a database. In the cases where the assessors have captured their responses using a tablet and Tangerine® software this is very easy and the data is available within minutes of the assessors uploading their data. In those cases where the assessors worked on a paper version of the test, the responses of the assessors for each subtest and item must be captured typically using Microsoft Excel.

Once all of the IRR data has been captured the score for each assessor needs to be calculated and a report needs to be produced. The scores are calculated (by subtest and for the EGMA overall) by comparing the response of the assessor with the response expected according to the Golden Standard and reporting the % of agreement. For example, for the number identification subtest there will be the responses for the 20 items, the time remaining, and the last item responded to in the time available. The time remaining is hard to get precisely and so it may be wise to allow a margin of error of up to 2 seconds either way from the Gold Standard. In terms of the last item responded to, if the assessor did not mark the same last item as the Golden Standard then all of their responses from that point on will be incorrect and they will be penalized in terms of the items. In summary each assessor's disagreements with the Golden Standard will be counted up, divided by 21 (20 items and the time remaining) and presented as a percentage. This percentage is the assessor's "percentage agreement" with the Golden Standard. Typically assessors should achieve a percentage agreement of more than 90% by the end of the training session.

EGMA overall	EGMA Number ID	EGMA Number disc.	EGMA Missing number	EGMA Add L1	EGMA Add L2	EGMA Sub L1	EGMA Sub L2	EGMA Word problems
92%	82%	78%	100%	100%	100%	100%	100%	77%
96%	100%	95%	80%	100%	97%	100%	100%	100%
95%	100%	100%	90%	98%	98%	80%	100%	92%
97%	100%	79%	100%	100%	95%	100%	100%	100%
92%	100%	93%	80%	88%	100%	80%	100%	92%
98%	100%	95%	100%	100%	91%	100%	100%	100%
88%	100%	90%	80%	94%	98%	60%	100%	85%
96%	100%	89%	90%	100%	95%	100%	96%	100%
93%	100%	99%	90%	100%	98%	60%	100%	100%
97%	100%	99%	80%	100%	98%	100%	96%	100%
98%	100%	92%	90%	100%	100%	100%	100%	100%
97%	100%	100%	100%	100%	97%	80%	100%	100%
96%	100%	100%	90%	100%	97%	80%	100%	100%
97%	100%	89%	90%	98%	98%	100%	100%	100%
95%	91%	89%	90%	98%	98%	100%	96%	100%
89%	91%	97%	90%	100%	97%	58%	100%	78%
92%	100%	100%	80%	98%	98%	60%	100%	100%
97%	100%	98%	100%	100%	98%	80%	100%	100%
97%	100%	100%	100%	100%	98%	80%	100%	100%

A typical report for an IRR test is shown above. The left hand column would typically have the assessor name. The different colors show the different levels of agreement: green indicates an agreement of 90% or more, orange an agreement of 75% or more and red an agreement of less than 75%. The report clearly indicates areas for improvement by each assessor and areas that need further training for all assessors.

4.4 Other

As we continue to implement the EGMA in new countries, we are learning and adjusting to the demands of host countries. In particular, we have developed new subtests and will be updating this toolkit shortly after we finalize the design of the tasks.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review, 2*, 213–236.

Baroody, A. J. (1987a). *Children's mathematical thinking*. New York: Teachers College Press.

Baroody, A. J. (1987b). The development of counting strategies for single-digit addition. *Journal for Research in Mathematics Education, 18*(2), 141–157.

Baroody, A. J. (2004). The developmental bases for early childhood number and operations standards. In D. H. Clements & J. Sarama (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics* (pp. 173–219). Mahwah, NJ: Lawrence Erlbaum Associates.

Baroody, A., Lai, M., & Mix, K. (2006). The development of young children's early number and operation sense and its implications for early childhood education. In B. Spodek & S. Olivia (Eds.), *Handbook of research on the education of young children* (pp. 187–221). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Blanton, M. L., & Kaput, J. J. (2004). Elementary students' capacity for foundational thinking. In M. J. Hoines & A. B. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology in Mathematics Education* (Vol. 2, pp. 135–142). Bergen, Norway.

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development, 70*, 1016–1031.

Carpenter, T. P., Fennema, E., & Franke, M. L. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *The Elementary School Journal, 97*(1), 3–20.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. (1999). *Children's mathematics: Cognitively guided instruction.* Portsmouth, NH: Heinemann, for the National Council of Teachers of Mathematics.

Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention, 30*(2), 3–14. http://dx.doi.org/10.1177/073724770503000202

Clarke, B., Baker, S., Smolkowski, K., & Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education, 29*, 46–57. http://dx.doi.org/10.1177/0741932507309694

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*(2), 234–248.

Clements, D. H., & Sarama, J. (2007). Early childhood mathematics learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 461–555). Charlotte, NC: Information Age Publishing.

Common Core Standards Writing Team. (2013). *Progressions for the Common Core State Standards in Mathematics (draft): K. Counting and Cardinality; K–5 Operations and Algebraic Thinking.* Tucson, AZ: Institute for Mathematics and Education, University of Arizona.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20,* 487–506.

De Smedt, B., Verschaffel, L., & Ghesquiere, P. (2009). The predictive value of numerical magnitude comparison for individual differences in mathematics achievement. *Journal of Experimental Child Psychology, 103,* 469–479.

Duncan, G. J., Claessens, A., Huston, A. C., Pagani, L. S., Engel, M., Sexton, H., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *Journal of Special Education, 41*, 121–139. http://dx.doi.org/10.1177/00224669070410020101

Fuchs, L. I., & Fuchs, D. F. (2004). Curriculum-based measurement: Describing competence, enhancing outcomes, evaluating treatment effects, and identifying treatment nonresponders. *Journal of Cognitive Education and Psychology, 4*, 112–130. http://dx.doi.org/10.1891/194589504787382929

Gay, J., & Cole, M. (1967). *The new mathematics and an old culture: A study of learning among the Kpelle of Liberia*. New York, NY: Holt, Rinehart, and Winston.

Geary, D. C. (1994). *Children's mathematical development*. Washington, DC: American Psychological Association.

Geary, D. C. (2000). From infancy to adulthood: The development of numerical abilities. *European Child & Adolescent Psychiatry, 9*, II11–II16.

Geary, D. C., Bow-Thomas, C. C., & Yao, Y. (1992). Counting knowledge and skill in cognitive addition: A comparison of normal and mathematically disabled children. *Journal of Experimental Child Psychology, 54*, 372–391.

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard, University Press.

Gersten, R., & Chard, D. J. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *Journal of Special Education, 33,* 18–28. http://dx.doi.org/10.1177/002246699903300102

Ginsburg, H. P., & Baron, J. (1993). Cognition: Young children's construction of mathematics. In R. J. Jensen (Ed.), *Research ideas for the classroom: Early childhood mathematics* (pp. 3–21). Reston, VA: National Council of Teachers of Mathematics.

Ginsburg, H. P., Klein, A., & Starkey, P. (1998). The development of children's mathematical thinking: Connecting research with practice. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology* (5th ed., Vol. 2: Cognition, perception, and language, pp. 401–468). New York: John Wiley & Sons.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development, 74*(3), 834–850.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45*, 850–867.

Klibanoff, R. S., Levine, S. C., Huttenlocher, J., Vasilyeva, M., & Hedges, L. (2006). Preschool children's mathematical knowledge: The effect of teacher "Math Talk." *Developmental Psychology, 42*(1), 59–69.

Langrall, C. W., Mooney, E. S., Nisbet, S., & Jones, G. A. (2008). Elementary students' access to powerful mathematical ideas. In L. D. English (Ed.), *Handbook of international research in mathematics education* (2nd ed., pp. 109–135). New York, NY: Lawrence Erlbaum Associates.

Linacre, J. M. (2013). *Winsteps® Rasch measurement computer program*. Beaverton, OR: Winsteps.com.

Malloy, C. E. (2008). Looking throughout the world for democratic access to mathematics. In L. D. English (Ed.), *Handbook of international research in mathematics education* (2nd ed., pp. 20-31). New York, NY: Lawrence Erlbaum Associates.

Martin, M. O., Mullis, I. V., & Foy, P. (2008). TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades. Boston, MA: IEA.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (NCTM). (2004b).

National Council of Teachers of Mathematics (NCTM). (2008).

Nunes, T., & Bryant, P. (1996). *Children doing mathematics*. Oxford, Great Britain: Blackwell.

Organisation for Economic Co-operation and Development (OECD). (2009). *Learning mathematics for life: A perspective from PISA.* Paris: OECD Publishing.

Perry, B., & Dockett, S. (2008). Young children's access to powerful mathematical ideas. In L. D. English (Ed.), *Handbook of international research in mathematics education* (2nd ed., pp. 75–108). New York, NY: Lawrence Erlbaum Associates.

Piazza, M., Pica, P., Izard, V., Spelke, E. S., & Dehaene, S. (2013). Education enhances the acuity of the nonverbal approximate number system. *Psychological Science, OnlineFirst version*, 1–7.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., & Jenkins, F. (2012). *Highlights from TIMSS 2011: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context* (NCES 2013-009 Revised). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Radford, L. (2008). Culture and cognition: Towards an anthropology of mathematical thinking. In L. D. English (Ed.), *Handbook of international research in mathematics education* (2nd ed., pp. 439–464). New York, NY: Lawrence Erlbaum Associates.

Reikeras, E. K. L. (2006). Performance in solving arithmetic problems: A comparison of children with different levels of achievement in mathematics and reading. *European Journal of Special Needs Education, 21*(3), 233–250.

Reubens, A., & Kline, T. (2009). *Pilot of the Early Grade Mathematics Assessment: Final report.* Prepared for USAID under the Education Data for Decision Making (EdData II) project, Task Order No. EHC-E-02-04-00004-00. Research Triangle Park, NC: RTI International. Retrieved from https://www.eddataglobal.org/math/index.cfm?fuseaction=pubDetail&ID=194

Romano, E., Babchishin, L., Pagani, L. S., & Kohen, D. (2010). School readiness and later acheivement: Replication and extension using a nationwide Canadian survey. *Developmental Psychology*, 46(5), 995-1007. doi: 10.1037/a0018880

RTI International. (2009a). *Early Grade Mathematics Assessment (EGMA): A conceptual framework based on mathematics skills development in children.* Prepared under the USAID Education Data for Decision Making (EdData II) project, Task Order No. EHC-E-02-04-00004-00. Retrieved from https://http://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=193

RTI International. (2009b). *Early Grade Reading Assessment toolkit.* Prepared for the World Bank, Office of Human Development, under Contract No. 7141961. Research Triangle Park, NC: Author. Retrieved from https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=149

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research*. New York, NY: Routledge.

Saxe, G. B. (1991). *Culture and cognitive development*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curricuum coherence: An examination of US mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, 37(5), 525-559. doi: 10.1080/0022027042000294682

Smith, E. (2008). Representational thinking as a framework for introducing functions in the elementary curriculum. In J. J. Kaput, D. W. Carraher & M. L. Blanton (Eds.), *Algebra in the early grades*. Mahwah, NJ: Lawrence Erlbaum Associates and National Council of Teachers of Mathematics.

Somerern, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The Think Aloud Method: A practical guide to modeling cognitive process.* San Diego, CA: Academic Press.

Starkey, P. (1992). The early development of numerical reasoning. *Cognition, 43*, 93–126.

Steen, L. A. (Ed.). (2001). *Mathematics and democracy: The case for quantitative literacy*. Washington, DC: National Council on Education and the Disciplines.

Tolar, T. D., Lederberg, A. R., & Fletcher, J. M. (2009). A structural model of algebra achievement: Computational fluency and spatial visualisation as mediators of the effect of working memory on algebra achievement. *Educational Psychology*, *29*(2), 239–266.

U.S. Department of Education. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Produced under U.S. Department of Education Contract No. ED04CO0082/0001 for Widmeyer Communications and No. ED04CO0015/0006 for Abt Associates, Inc. Washington, DC: Author. Retrieved from http://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

Vygotsky, L. (1978). *Mind in Society: The Development of Higher Psychological Processes* (M. Cole, V. John-Steiner, S. Scribner & E. Souberman, Trans.). Cambridge, MA: Harvard University Press.

Appendix A

Assessor Observation Checklist (EGMA)

Supervisor: _____ Assessor: _____Date: _____

	Checklist Items	Observed?
	GENERAL INSTRUCTIONS: DEVELOPING RAPPORT; GAINING CONSENT	✓, ✗ or n/a
	1. Assessor is relaxed and makes the child feel comfortable.	
**	2. Assessor reads aloud the consent text verbatim, obtains a verbal response, and checks the verbal consent box. If child declines to participate, assessor thanks the child and tells him/her to return to class.	
	3. Assessor completes data on first page, including the time test is started.	
	TASK 1: NUMBER IDENTIFICATION (Timed)	
	4. Assessor opens stimulus book to the correct page.	
**	5. Assessor follows script of instructions to the child, without adding unnecessary words.	
**	6. Assessor uses stopwatch correctly: Setting the countdown timer to 1 min (60 sec), starting when the child first speaks, stopping at the end of the items or after 60 seconds have passed.	
**	7. Assessor marks incorrect responses with a slash through the middle of the item.	
**	8. If child hesitates for 5 seconds, assessor points to the next item and says, “Go on” and marks the item as incorrect	
**	9. Assessor marks a bracket at the point reached by child after 60 seconds have passed and enters 0 seconds remaining. If child completes task before 60 seconds have passed, the assessor stops the stopwatch, places a bracket after the last item, and enters the seconds remaining.	
	10. Assessor holds response sheet on clipboard outside of child’s visual range.	
	TASK 2: NUMBER DISCRIMINATION (Not timed)	
	11. Assessor conducts practice items: • Opens stimulus book to the correct page. • Follows script of instructions to the child, without adding unnecessary words.	
	12. Assessor opens stimulus book to the correct page	
**	13. Assessor follows script of instructions to the child, without adding unnecessary words.	
**	14. Assessor makes sure that the child names the larger number and does not merely point at the number.	
**	15. Assessor checks the appropriate response (0 = incorrect and 1 = correct) for each item	
**	16. If child makes 4 consecutive errors the assessor stops the child and moves to the next task.	
	17. If child hesitates for 5 seconds, assessor points to the next item and says, “Go on” and marks the item as incorrect	
	18. Assessor uses a blank page to cover the items on the stimulus page and manage the child’s progress (if necessary)	
	19. Assessor holds response sheet on clipboard outside of child’s visual range.	

	Checklist Items	Observed?
	TASK 3: MISSING NUMBER (Not timed)	
	20. Assessor conducts practice items: • Opens stimulus book to the correct page. • Follows script of instructions to the child, without adding unnecessary words.	
	21. Assessor opens stimulus book to the correct page	
**	22. Assessor follows script of instructions to the child, without adding unnecessary words.	
**	23. Assessor checks the appropriate response (0 = incorrect and 1 = correct) for each item	
**	24. If child makes 4 consecutive errors the assessor stops the child and moves to the next task.	
	25. If child hesitates for 5 seconds, assessor points to the next item and says, "Go on" and marks the item as incorrect	
	26. Assessor uses a blank page to cover the items on the stimulus page and manage the child's progress (if necessary)	
	27. Assessor holds response sheet on clipboard outside of child's visual range.	
	TASK 4A: ADDITION LEVEL 1 (Timed)	
	28. Assessor opens stimulus book to the correct page.	
**	29. Assessor follows script of instructions to the child, without adding unnecessary words.	
**	30. Assessor uses stopwatch correctly: Setting the countdown timer to 1 min (60 sec), starting when the child first speaks, stopping at the end of the items or after 60 seconds have passed.	
	31. If the child repeats the whole problem the assessor tells the child that it is not necessary to do so and that it is enough to give the correct answer only.	
**	32. Assessor marks incorrect responses with a slash through the middle of the item.	
**	33. If child hesitates for 5 seconds, assessor points to the next item and says, "Go on" and marks the item as incorrect	
**	34. Assessor marks a bracket at the point reached by child after 60 seconds have passed and enters 0 seconds remaining. If child completes task before 60 seconds have passed, the assessor stops the stopwatch, places a bracket after the last item, and enters the seconds remaining.	
	35. Assessor uses a blank page to cover the items on the stimulus page and manage the child's progress (if necessary)	
	36. Assessor holds response sheet on clipboard outside of child's visual range.	
	TASK 4B: ADDITION LEVEL 2 (Not timed)	
	37. Assessor opens stimulus book to the correct page.	
**	38. Assessor follows script of instructions to the child, without adding unnecessary words.	
**	39. Assessor makes it clear to the child that they may use the paper and pencil to solve the problems but that they do not have to.	
	40. If the child repeats the whole problem the assessor tells the child that it is not necessary to do so and that it is enough to give the correct answer only.	
**	41. If the child uses an inefficient strategy the assessor asks the child if they "know another way of doing the calculation" if the child continues to use an inefficient strategy the assessor points to the next item and says, "Go on" and marks the item as incorrect.	
**	42. Assessor checks the appropriate response (0 = incorrect and 1 = correct) for each item	
**	43. If child hesitates for 5 seconds, assessor points to the next item and says, "Go on" and marks the item as incorrect	
**	44. Assessor completes at least one check box to describe how the child performed the calculation	

	Checklist Items	Observed?
	45. Assessor uses a blank page to cover the items on the stimulus page and manage the child's progress (if necessary)	
	46. Assessor holds response sheet on clipboard outside of child's visual range.	
	TASK 5A: SUBTRACTION LEVEL 1 (Timed)	
	47. Assessor opens stimulus book to the correct page.	
**	48. Assessor follows script of instructions to the child, without adding unnecessary words.	
**	49. Assessor uses stopwatch correctly: Setting the countdown timer to 1 min (60 sec), starting when the child first speaks, stopping at the end of the items or after 60 seconds have passed.	
	50. If the child repeats the whole problem the assessor tells the child that it is not necessary to do so and that it is enough to give the correct answer only.	
**	51. Assessor marks incorrect responses with a slash through the middle of the item.	
**	52. If child hesitates for 5 seconds, assessor points to the next item and says, "Go on" and marks the item as incorrect	
**	53. Assessor marks a bracket at the point reached by child after 60 seconds have passed and enters 0 seconds remaining. If child completes task before 60 seconds have passed, the assessor stops the stopwatch, places a bracket after the last item, and enters the seconds remaining.	
	54. Assessor uses a blank page to cover the items on the stimulus page and manage the child's progress (if necessary)	
	55. Assessor holds response sheet on clipboard outside of child's visual range.	
	TASK 5B: SUBTRACTION LEVEL 2 (Not timed)	
	56. Assessor opens stimulus book to the correct page.	
**	57. Assessor follows script of instructions to the child, without adding unnecessary words.	
**	58. Assessor makes it clear to the child that they may use the paper and pencil to solve the problems but that they do not have to.	
	59. If the child repeats the whole problem the assessor tells the child that it is not necessary to do so and that it is enough to give the correct answer only.	
**	60. If the child uses an inefficient strategy the assessor asks the child if they "know another way of doing the calculation" if the child continues to use an inefficient strategy the assessor points to the next item and says, "Go on" and marks the item as incorrect.	
**	61. Assessor checks the appropriate response (0 = incorrect and 1 = correct) for each item	
**	62. If child hesitates for 5 seconds, assessor points to the next item and says, "Go on" and marks the item as incorrect	
**	63. Assessor completes at least one check box to describe how the child performed the calculation	
	64. Assessor uses a blank page to cover the items on the stimulus page and manage the child's progress (if necessary)	
	65. Assessor holds response sheet on clipboard outside of child's visual range.	
	TASK 6 WORD PROBLEMS (Not timed)	
	66. Assessor conducts practice item: • Follows script of instructions to the child, without adding unnecessary words.	
**	67. Assessor follows script of instructions to the child, without adding unnecessary words.	
**	68. Assessor makes it clear to the child that they may use the counters, paper and pencil to solve the problems but that they do not have to.	
**	69. Assessor takes care to "pause and check" after each phrase in the problem (as per the instructions) making sure that the child is engaged and interested in the problem	
**	70. Assessor checks the appropriate response (0 = incorrect and 1 = correct) for each item	

	Checklist Items	Observed?
**	71. If child makes 4 consecutive errors the assessor stops the child and moves to the next task.	
	72. If child makes no attempt to solve the problem for 5 seconds, assessor says, "let us try another one," starts the next problem and marks the item as incorrect.	
**	73. If child has worked on the problem for more than 60 seconds and not produced an answer, assessor says, "let us try another one," starts the next problem and marks the item as incorrect.	
	74. Assessor holds response sheet on clipboard outside of child's visual range.	
	INTERVIEW QUESTIONS	
	75. Assessor maintains an appropriate pace through the interview – neither taking too long nor rushing the child. The atmosphere is one of a conversation	
	76. Assessor takes care NOT to read the responses for the items where the child is asked to respond freely (items 5, 14, 15, 18, 20 and 21)	
	77. Assessor observes the skip rules correctly where appropriate (items 8/9, 11/12/13, 16/17, 19/20)	
	78. Assessor completes the responses correctly for each item including: • Multiple responses where appropriate (items 5, 14, 15, 18, 20 and 21)	
	79. Assessor thanks the child appropriately on completion of the interview reminding the child not to discuss the experience with the other children.	
	80. Assessor gives the child a thank you gift	
	81. Assessor records the time completed	

****MAJOR ERRORS**

Notes: