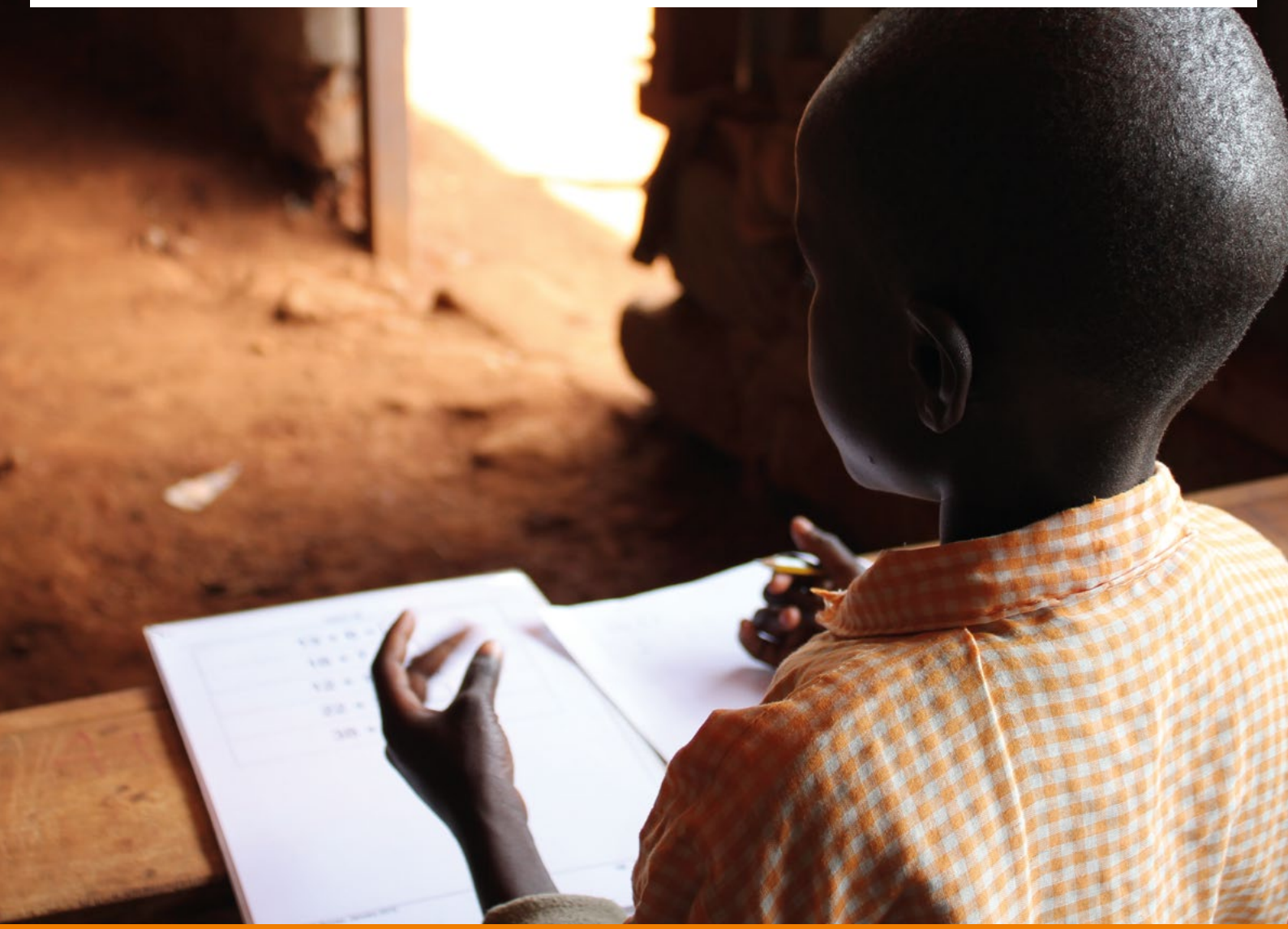
Federal Ministry
for Economic Cooperation
and Development



Assessment tools

Desk Study on Assessment Tools for Numeracy Education in Pre-School and Early Grades

Published by giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

Federal Ministry
for Economic Cooperation
and Development

Assessment tools

Desk Study on Assessment Tools for Numeracy Education in Pre-School and Early Grades

Published by

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

Acknowledgement

The author, Jeff Davis, would like to thank the German Federal Ministry for Economic Cooperation and Development (BMZ) and the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH for their continued support during the study.

In particular, thanks go to the following people:

Dr. Michael Holländer
Senior Advisor Education
Sector Programme Numeracy Education Section
Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Lena Mächel
Junior Advisor Education
Sector Programme Numeracy
Education Section
Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Portions of this study were presented at the BMZ/GIZ/GPE/USAID All Children Learning (ACL) Conference in Rabat, Morocco on December 02-05, 2013 and at the Comparative and International Education Society (CIES) Annual Conference in Toronto, Canada on March 10-15, 2014.

This study was commissioned by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) Education Section on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ). The analysis, results and recommendations in this paper represent the opinion of the author and are not necessarily representative of the position of the GIZ and BMZ.

List of abbreviations

ACER	Australian Council for Educational Research
ACL	All Children Learning
AERA	American Educational Research Association
APA	American Psychological Association
ASER	Annual Status of Education Report
BMZ	German Federal Ministry for Economic Cooperation and Development
CIES	Comparative and International Education Society
CUE	Center for Universal Education (Brookings Institution)
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
EDA	Education Development Associates LLC
EFA	Education for All
EGMA	Early Grades Mathematics Assessment
EGRA	Early Grades Reading Assessment
EMDA	Early Math Diagnostic Assessment
ENT	Early Numeracy Test
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH
GPE	Global Partnership for Education
ICDM	I Can Do Maths
IEA	International Association for the Evaluation of Educational Achievement
IQ	Intelligence Quotient
IRT	Item Response Theory
ITC	International Test Commission
KeyMath	KeyMath Diagnostic Assessment
LMTF	Learning Metrics Task Force
MMY	Mental Measurements Yearbook
NCME	National Council for Measurement in Education
NCTM	National Council of Teachers of Mathematics
NGO	Non-Governmental Organization
NMAP	National Mathematics Advisory Panel
OLO	Observatory of Learning Outcomes
PAL II	Process Assessment of the Learner II – Math
PASEC	Programme d'Analyse des Systèmes Educatifs de la CONFEMEN
PIPS	Performance Indicators in Primary School
PIRLS	Progress in International Reading and Literacy Study
SAFED	South Asian Forum for Educational Development
SACMEQ	Southern Africa Consortium for Monitoring Educational Quality
SEA	School Entry Assessment
TEAM	Tools for Early Assessment in Math
TEMA	Test of Early Mathematics Ability
TIMSS	Trends in International Mathematics and Science Study
UIS	UNESCO Institute of Statistics
USAID	U.S. Agency for International Development
Uwezo	"Capability" in Kiswahili

Table of contents

Abstract

In September 2013, the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ) commissioned a desk study in response to needs by education officials, development practitioners, donor representatives, researchers, and academics for information on numeracy assessment tools for pre-school and the early grades for use in low-income countries. It was a follow-up to a 2012 desk study, also commissioned by the BMZ/GIZ, which provided information on the basis for assessment tools for numeracy education in developing countries. From that previous study, requests were made to 1) develop a user-friendly format for classifying assessment tools that could be adapted to a database; and 2) identify specific assessment tools, review them, and adapt them to the format. The study was conducted from September 2013 to May 2014. Both of the stated goals were achieved. Parts of the study were presented at international conferences in Rabat, Morocco (2013) and Toronto, Canada (2014), and feedback was gathered and incorporated into the present document.

Executive Summary

In September 2013, the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ) commissioned this desk study on assessment tools for numeracy education in pre-school and early grades. It was a follow-up to a 2012 desk study, also commissioned by the BMZ/GIZ, that described the basis for developing numeracy assessment tools for use in socio-economically less developed countries. Both studies are a part of GIZ's Sector Programme Numeracy that supports the promotion of children's acquisition of basic numeracy competencies, such as measuring, estimating, and operations. The purpose of the two studies was to provide information on assessment and assessment tools that could help improve learning outcomes in numeracy. The studies support the Education for All (EFA) goals of "recognized and measureable learning outcomes (that) are achieved by all, especially in literacy, numeracy, and essential life skills" as well as objective 3 of GPE's strategic plan for "a dramatic increase occurs in the number of children who are learning and demonstrating mastery of basic literacy and numeracy skills by grade 3."

From September 2013 to May 2014, the desk study was conducted for a target audience of practitioners, development workers, donor representatives, researchers, and academics who needed information on numeracy assessment tools. Two main goals of the study were identified in the terms of reference: 1) develop a user-friendly format for classifying early mathematics assessment tools; and 2) identify and review tools for use in early mathematics assessment in developing countries. Parts of the study were presented at two conferences – in Rabat, Morocco (2013) and Toronto, Canada (2014) – from which feedback was solicited for the final version of the study.

For the tools review, the landscape was surveyed through conversations with mathematics experts and a web search. Ten tools were identified for initial review and formatting:

1. Annual Status of Education Report - Mathematics (ASER)
2. "Capability" in Kiswahili – Mathematics (Uwezo)
3. Early Grade Math Assessment (EGMA)
4. Early Numeracy Test (ENT)
5. I Can Do Maths (ICDM)
6. KeyMath Diagnostic Assessment (KeyMath-3)
7. Test of Early Mathematics Ability (TEMA-3)
8. Tools for Early Assessment in Math (TEAM)
9. Trends in International Mathematics and Science Study (TIMSS)
10. TIMSS Numeracy

First, information on each assessment was described in five categories: needs addressed, test development, competencies measured, administration/scoring/reporting, and strengths/weaknesses. Second, a classification scheme (or format) was developed that included several categories of assessment information: overview, content, implementation, analysis/reporting, recommendations, further information, and user feedback. In addition, linkages are made between the tools and the content domains identified by the Learning Metrics Task Force (LMTF), which was sponsored by the UNESCO Institute of Statistics (UIS) and the Brookings Institution Center for Universal Education (CUE).

The desk study is considered as a starting point for a user-friendly database of numeracy assessment tools.

Introduction

In September 2013, the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ) and the commissioned Education Development Associates LLC (EDA) conducted a desk study on assessment tools for numeracy education in pre-school and early grades. It is a follow-up to a previous BMZ/GIZ-sponsored study on the assessment of learning outcomes with reference to early grade numeracy in low-income countries (GIZ, 2012).

Both studies have been conducted in support of GIZ's Sector Programme Numeracy to promote mathematics competencies in pre-school and in the early grades (GIZ website, 2013). GIZ believes that basic numeracy competencies – such as measuring, estimating, and operations – form the "foundation for all future learning and provide opportunities for children to become active members of society." This is in accordance with the 2015 Education for All (EFA) goals, for which "recognized and measureable learning outcomes are achieved by all, especially in literacy, numeracy, and essential life skills."

BMZ/GIZ has been an active supporter of the EFA goals for example by: 1) sponsoring a series of technical papers on early numeracy topics; 2) co-sponsoring conferences in Germany (December 2012) and Morocco (December 2013) to promote better instructional practices and improved learning outcomes in numeracy; and, 3) leading an early numeracy community of practice, which had been initially formed through the Global Partnership for Education (GPE). Together, these three activities support GIZ's four areas for fast tracking EFA's numeracy goals for children:

1. Numeracy skills development in pre-school and early grades
2. Learning outcomes and assessments in numeracy
3. Mobile education for numeracy
4. Utilizing synergies and lessons learned from literacy.

As with the 2012 study, the present study supports GIZ's second area for children's numeracy development. In the terms of reference, GIZ stated that it is a "preparatory step for the launch of a user-friendly and accessible database for assessment tools in numeracy education." The users of the database will include practitioners who need assessment tools to measure the learning outcomes of students for the improvement of education quality in numeracy. The assessment tools, which are only comprised of summative assessments at this point, will be used to set baselines, monitor progress towards learning goals, provide information for improved instruction, and guide political dialogue.

The target group for the tools includes those that are appropriate for pre-school to early grades in developing countries. Tools that have been developed for transitioning countries may also be taken into consideration if they have the possibility of adaptation to developing country contexts. In an effort to help improve the assessment of early numeracy, the study has two main objectives:

1. Develop a format for presentation of information on assessment tools. The format needs to allow for browsing against selection criteria (categories, y/n, short answer) as well as the option for obtaining in-depth information (description).
2. Provide a comprehensive review of the landscape of numeracy assessment tools in text format, both through descriptions and classifications. It will focus on key characteristics of the tools and recommendations for the use of the tools in given situations.

The study will specify the initial structure of the database, and it will also give background information on various assessment tools for populating the initial database.

Methodology of Data Collection

All of the well-known international assessments of early numeracy in developing countries (e.g., EGMA, ASER, Uwezo, TIMSS, TEMA, etc.) are included in the review. Selected lesser-known assessment tools – but with potential as contributors to international best practices – are included as well. Some of these lesser-known tools were identified through internet searches. Others were identified through the Buros Center for Testing, which publishes descriptions and critical evaluations of commercially available tools through its *Test Reviews Online, Tests in Print, and Mental Measurements Yearbook (MMY)*.

After selecting tests that are either used internationally or that have potential for international use, they were described with information from the following categories:

1. Needs addressed
2. Test development
3. Competencies measured
4. Administration/scoring/reporting
5. Strengths/weaknesses

Then, the descriptive information from those tests was catalogued according to evaluation criteria that were developed for the purposes of this study (see below). The idea from these criteria is to form the basis for providing information on each assessment in a database. This information is presented in the annex.

Three limitations of the initial data collection are particularly important.

1. The first limitation is that the tools are limited to English-language versions. Many of the tools have been administered in other languages, but there needed to be an English-language version for inclusion in this report due to the author's language background.
2. The second limitation is that the tools have different levels of accessibility. Some of the tools – such as EGRA, ASER, and Uwezo – are open source. Other tools – such as TEMA-3, TEAM, KeyMath-3, and ICDM – are commercial products, with costs up to a few hundred dollars. Another tool reviewed – ENT – is available from researchers in different countries. Finally, the TIMSS tools are secure assessments, so that only some of the items are accessible to the general public.
3. The third limitation is that the study does not cover national or regional assessments. Many countries are now conducting national assessments using their own internally developed tools. In general, these tools are not available for public use. In addition, there are also regional assessments, such as those by the Southern Africa Consortium for Monitoring Educational Quality (SACMEQ) and the Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC), for which tools are also not publically available. The UNESCO Institute of Statistics (UIS), under the Observatory of Learning Outcomes (OLO) initiative, is conducting a review of these kinds of tools.

Selection of Evaluation Criteria

The first objective for the study involves developing evaluation criteria and a format for presenting information on assessment tools. Information for this section was gathered from the following sources:

1. International Guidelines for Test Use by the International Test Commission (ITC)
2. Standards for Educational and Psychological Testing by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council for Measurement in Education (NCME)

Based on this review, the following list includes a classification scheme for the criteria and characteristics of the assessments. The scheme has areas, sub-areas, and categories (multiple, yes/no, description). The organization of the scheme is intended to provide the user with information on each assessment in a consistent format.

1. Overview
    a. Type: formative, diagnostic, examination, assessment (national/international)
    b. Target group: pre-school, school entry, grades 1-4
    c. Administration: individual (oral), group (paper and pencil)
    d. Developer: government, non-governmental organization (NGO), donor agency, vendor (and name of developer)
    e. Accessibility: open, closed
    f. Cost structure: free, purchase (and how much)
    g. Countries: list
2. Content
    a. Learning domains: counting, number sense, operations, problem solving
    b. Cognitive domains: knowledge, comprehension, application
    c. Item type: short answer, objective, multiple choice, open-ended
    d. Number of items: total and by learning domain
    e. Adaptation to local context: yes, no
    f. Alternate forms: A, B, etc.
3. Implementation
    a. Materials: guides, record-keeping forms, manipulatives, CD
    b. Administration time: per student or class
    c. Enumerator training: administration guide, video, webinar
    d. Timed items: yes, no
    e. Technology-based application: mobile, tablet, laptop

4. Analysis and reporting
    a. Scoring: raw scores, item scores
    b. Reliability: coefficient alpha, test-retest
    c. Score reports: yes, no (description)
    d. Scaling: yes, no (description)
    e. Performance categories: yes, no (levels)
    f. Decision tree: process for determining level based on scores
    g. Performance level descriptors: yes, no (description)
    h. Computer application: yes, no
    i. Timing: typical amount time needed for the analysis and reporting

In addition to the overview, content, implementation, and analysis/reporting, there are also brief sections on recommendations, further information, and user feedback.

5. Recommendations
    a. Strengths (description)
    b. Weaknesses (description)
    c. Suggestions (description)
6. Further information
    a. Assistance: tool developer, support group
    b. Point of contact: person, email address, telephone number
    c. Link: website
7. User feedback
    a. Rating system (five stars)
    b. Comments

The descriptive summaries of each test are presented below. More detailed and quantifiable information on the tests by category is provided in the appendix. Perhaps the least defined characteristic of the assessments is their cost. The commercial products require an initial outlay, but they may end up being less expensive to develop or adapt due to the work that was carried out by the test developers and publishers to prepare the tests. In addition, by far the majority of the costs are in the field administration of the tests, so the processes of development/adaptation and scoring/analysis/reporting are relatively low percentages of the total budget. The exception is the ASER and Uwezo tests, which are administered by village volunteers and therefore have low field costs. Organizations that have developed/adapted, administered, and scored/analyzed/reported the results, as well as donors that have funded such efforts, may be willing to share their budgets and/or associated costs; however, it was beyond the scope of this study to gather and examine such information.

Descriptive Information

Based on an international review of the assessment landscape, descriptive information is provided on the following diagnostic assessment tools (in alphabetical order):

1. Annual Status of Education Report - Mathematics (ASER)
2. "Capability" in Kiswahili – Mathematics (Uwezo)
3. Early Grade Math Assessment (EGMA)
4. Early Numeracy Test (ENT)
5. I Can Do Maths (ICDM)
6. KeyMath Diagnostic Assessment (KeyMath-3)
7. Test of Early Mathematics Ability (TEMA-3)
8. Tools for Early Assessment in Math (TEAM)

In addition, two summative assessments were reviewed due to their international implementation. However, the tests themselves are not available (only sample questions) and countries need to make a request to the International Association for the Evaluation of Educational Achievement (IEA) for their administration.

9. Trends in International Mathematics and Science Study (TIMSS)
10. TIMSS Numeracy

The ten assessments are described below using the five categories.

Annual Status of Education Report – Mathematics (ASER)

1. **Needs Addressed:**
   The ASER center in India, under Pratham (an NGO), developed a household-based, locally administered national-level assessment in basic literacy and numeracy (grades 1 to 4). There is also now an ASER center in Pakistan in association with the South Asian Forum for Educational Development (SAFED). ASER is currently the most comprehensive data source on elementary school children's learning levels in the two countries. These assessments were developed out of a need to provide independent evidence of children's progress towards learning basic literacy and numeracy skills. The ASER assessments are not designed for use in schools (they are part of household surveys) but the information could be used to help in planning for education programs.

2. **Test Development:**
   ASER began to collect data in 2005 using internally developed tools. The India tools are aligned to the content in grades 1-4 of the National Curriculum Framework. The Pakistan tools are similar, with alignment to the new Pakistani curriculum (though the curriculum has not been uniformly implemented in all provinces). The assessment items are designed for individual administration using a handbook for village volunteers. The measures have been validated through several studies. The ASER study also includes information on infrastructure, school enrolment, and attendance.

3. **Competencies Measured:**
   The ASER assessment has used a core set of tasks from 2005 to 2011, which were given to all children between the ages of 5-16. There have been other tasks that were used infrequently. For example, in 2006, ASER used word problem items involving subtraction and division. In 2007 and 2008, there were word problems with currency and telling time. In 2010, there were calendar, area, and estimation tasks. The core tasks given every year are:

- Number Recognition: 1-9
- Number Recognition: 11-99
- Subtraction: 2-digit numbers with borrowing
- Division: 3-digit number by a 1-digit number with remainders

4. **Administration/Scoring/Reporting**
ASER recruits a large number of volunteers from NGOs, citizen groups, and educational and government institutions to assess children each year. ASER provides a short document explaining the assessment procedures. For example, the assessor is told to begin with the subtraction problem; if the child gets two incorrect responses in a row, the assessor moves to number recognition tasks 11-99. For each set of items (i.e., subtraction problems), the assessor determines if an answer is correct or incorrect, which then determines the subsequent problems given to the child. If a child solves a certain number of problems in a set correctly, he/she is scored as "a child that can do subtraction." If a child cannot solve a certain number of problems correctly, he/she is scored as "a child that cannot do subtraction." The assessor scores each assessment as it is being done. Then, community members gather and create a village report card, which has information on enrollment, attendance, and basic learning for all children in the village. Based on this information, next steps are planned by the community.

5. **Strengths/Weaknesses**
Strengths are that a) the assessment tools are easy to use by volunteers with minimal training; b) the results are generated quickly and are easy to interpret, and can provide immediate feedback to teachers, parents, and other invested in education; c) the assessments are easily scalable (they currently assess about 700,000 children annually in India); and d) it encourages parents/local community members to take responsibility for children's learning. Weaknesses are that a) the ASER tools measure largely procedural knowledge, without focusing on conceptual knowledge; b) the tools also focus largely on formal mathematics with little informal mathematics; c) there may also be problems with standardized implementation due to scale and use of minimally trained volunteers to assess children; and d) a lack of psychometric information.

"Capability" in Kiswahili (Uwezo) - Mathematics

1. **Needs Addressed**
The Uwezo assessments were modeled after ASER. They address a need for early literacy and numeracy in East Africa, namely in Kenya, Tanzania, and Uganda. The headquarters for East Africa is in Nairobi, Kenya. It is the main independent data source on elementary school children's learning levels in those three countries. Each year, Uwezo produces three national reports along with one regional report with a summary of the findings across the three countries. Their philosophy is to trigger improvements in children's learning by a) conducting countrywide annual assessments; b) communicate the findings widely and encourage debate; c) shift the focus from schooling inputs to outcomes; d) learn from experience and make adjustments each year.

2. **Test Development**
Uwezo began to collect data in 2009 using tools that were derived from the ASER assessments. The tools are aligned to the Grade 2 curriculums in each country, and thus are somewhat different in each place. They are individually administered in households. Along with the tests data, other information is collected on age, pre-school education, parental education, class size, attendance, teachers, learning environment, and facilities.

3. **Competencies Measured**
The Uwezo assessments have used a core set of tasks from 2009 to 2013 for children between the ages of 6-16. There have been other tasks included in some of the assessments, such as shapes and time. The core tasks developed in 2009 are:

- Counting Objects: 1-9
- Number Recognition: 11-99
- Place Value: Ones, tens, hundreds
- Addition: 2- and 3-digit numbers without carrying
- Subtraction: 2- and 3-digit numbers without borrowing
- Multiplication: 1-digit facts
- Division: 1- and 2-digit facts

4. **Administration/Scoring/Reporting**
Similarly to ASER, Uwezo recruits a large number of volunteers from NGOs, citizen groups, and educational and government institutions to assess children each year. Uwezo provides a short document explaining the assessment procedures. Testing begins with subtraction, and then goes forward (to multiplication) or backward (to addition) depending on the child's responses. For example, the assessor is told to begin with the subtraction problem; if the child gets two incorrect responses in a row, the assessor moves in reverse order of difficulty to addition, followed by place value, number recognition, and counting, again depending on whether the student answers the tasks correctly. If the child answers two subtraction problems correctly, the assessor moves to multiplication and then, if successful, to division. The child's results are given to the family after testing, and then submitted to the central office for processing. The central office in each country produces annual reports, ideally within 100 days of the testing.

5. **Strengths/Weaknesses**
As with ASER, the strengths of Uwezo assessment tools are a) ease of use by volunteers with minimal training; b) quickly generated and easily interpretable results; c) immediate feedback to teachers, parents, and other invested in education; d) ease of scalability due to manner of implementation and low cost (they currently assess about 350,000 children annually); and e) encouragement of parents/local community members to take responsibility for children's learning. In terms of weaknesses, the Uwezo tools have a) limited content coverage and measure largely procedural knowledge, without focusing on conceptual knowledge; b) a focus on formal mathematics instead of a balance with informal mathematics; c) potential problems with standardization due to the large numbers of participants and the use of minimally trained volunteers to assess children; and d) a lack of psychometric information.

Early Grade Math Assessment (EGMA)

1. **Needs Addressed**
EGMA was created started in 2010 by a team of consultants under the leadership of RTI International with funding from USAID and the World Bank. It was designed to address the need for an early mathematics diagnostic assessment (generally grades 1 to 3) in less developed countries. EGMA was a follow up to the success of Early Grades Reading Assessment (EGRA), which was developed starting in 2005.

2. **Test Development: Steps in the EGMA development process included the following**
a) an extensive literature review that established an assessment framework; b) the formation of a group of early mathematics experts to discuss skills and tasks; c) the development of a draft instrument, followed by piloting and revisions; d) the implementation of the instrument, including adaptations for different contexts; and e) the re-convening of the expert panel to review the implementation and make recommendations for refining the tasks and instructions.

3. **Competencies Measured**
Competencies covered by EGMA include both conceptual understanding and skills (i.e., procedural fluency/automaticity). There is more focus on formal (school, or symbolic) mathematics and less focus on informal (out-of-school, or non-symbolic) mathematics. Items such as oral counting, one-to-one correspondence, and number lines have been used in the past but are not included in the most recent version of the test. The core EGMA instrument has six tasks, each of which has multiple parts. Three of the sections are timed, e.g., the child has one minute to complete the task. The tasks are the following:

   - Number Identification (Timed): Reading numbers of 1- to 3-digits.
   - Number Discrimination: Stating which number is greater in value.
   - Missing Numbers: Filling in missing numbers by using patterns.
   - Addition Problems (Timed): Adding 1- and 2-digit numbers.
   - Subtraction Problems (Timed): Subtracting of 1- and 2-digit numbers.
   - Word Problems: Solving word problems involving the four basic operations

   In addition, two tasks are in the process of being developed for EGMA:

- Relational Reasoning: Using logic, such as the commutative property, to solve addition and subtraction problems.

- Spatial Reasoning: Visualizing spatial patterns and objects, and also mentally manipulating them.

4. **Administration/Scoring/Reporting**
The instructions for EGMA begin with telling the child about the assessment and trying to make the child feel at ease. The assessor then fills in information about the child on the instrument and begins with the items. Each item has its own set of instructions. The assessor stops the administration of the item if the child answers a successive number of parts incorrectly or if the child delays for more than a few seconds. If the item is timed, the assessor needs to use a stopwatch. On some of the items, there is a practice item so that the child has a better idea of what they are supposed to do. Each item is scored using a standard protocol. The assessor notes the item scores on a score sheet. Scores are reported in terms of number correct and number correct out of one minute (for the timed items). Some analysts have created percentage correct scores and grand means for a total score on the assessment. All scores are criterion-referenced.

5. **Strengths/Weaknesses**
Strengths of the EGMA instrument include a) high face validity, with universal tasks that represent a progression of foundational skills; b) the possibility for context-specific adaptations; c) measurement of both concepts and skills; d) better reliability through oral, one-to-one administration; and e) proven ability to "measure the pulse" of the general level of children in early mathematics in less developed countries. Weaknesses of the test include a) relatively high costs for administration, since only about 20 children can be assessed in one day by a test administrator; b) improbability of being able to collect data on each student, which would be useful for diagnostic assessment; c) lack of development on a country-specific basis, so alignment to the curriculum for individual countries must be done on a post hoc basis; d) lack of familiarity by most countries with oral administration, so initial discussions with education officials and training needs can be extensive; and e) issues with sustainability (in light of the other weaknesses).

Early Numeracy Test (ENT)

1. **Needs Addressed**
The first version of the ENT was developed in the Netherlands as the Dutch Early Numeracy Test in 1994. It was adjusted for use with young Finnish children (the Finnish Early Numeracy Test) in 2006. It is based on a developmental view of children's early numeracy, particularly as defined by Piaget (1965) and Ginsburg (2004). It focuses on several aspects of numerical and non-numerical knowledge. The ENT is valid for children in pre-school and early elementary school (ages 4 to 8).

2. **Test Development**
The ENT was based on research literature and existing teaching materials in the Netherlands. It went through several versions, including the Utrecht Test for Number Sense, the Early Mathematical Competence Test, and finally an expert panel evaluated the Early Numeracy Test. An expert panel in the Netherlands evaluated items for inclusion in the final version of the ENT. The items were piloted, and two forms of the test (A and B) were created with 40 items each. There was considerable research into the validity and reliability of the test, again in both the Netherlands and Finland. There were also predictive studies showing how the preschoolers' performance on the ENT predicted their performance in early elementary school. The test also went through a norming process in both the Netherlands and Finland, including with children from ethnic minorities, multi-language backgrounds, and those with special educational needs. In addition, the test was used in several other European countries (Germany, Belgium, Greece, England, Spain, and Slovenia) as well as in Hong Kong and Singapore.

3. **Competencies Measured**
The ENT is an individually administered tool that takes about 30 minutes per child. The content domains are the following:

- Knowledge of Quantity

- Concepts of Comparison

- Classification

- One-to-One Correspondence

- Seriation (Sorting Objects by Size or Shape)

- Number Words
- Structured and Resultative Counting
- Understanding of Numbers

There are a total of 40 items. The first 20 items are based on the logical principles underlying children's understanding of quantities and relations. The last 20 items focus more explicitly on the use and understanding of whole numbers. Factor analyses have resulted in classifying the items into two sub-domains (mathematical prerequisites, counting skills) or three sub-domains (mathematical prerequisites, counting skills, and general knowledge of numbers). The test is not timed.

4. **Administration/Scoring/Reporting**
The ENT has instructions for each item and is intended for use by trained assessors. Test materials – pictures, cubes, and paper-and-pencil) are provided to the assessors. The process involved administering each of the items to the children. Each item is worth one point and scored as either right or wrong. The assessor calculates a total score for each child. The test does not have timed components, though a child is limited on the amount of time to complete a task. Feedback is not provided to the child on whether they answer correctly or not. Reporting is based on the total score, subdomain scores (items 1-20 and 21-40), and individual items. Norms from the Netherlands and Finland are also used for the score interpretation. Information is collected on the child's age, gender, mother's education level, father's education level, number of children in the family, birth order of the child, and the child's hand preference.

5. **Strengths/Weaknesses**
The main strengths of the ENT are a) a basis on extensive research into children's early mathematics learning; b) experience with the test in different countries and with various languages; c) testing both concepts and skills; d) establishing strong validity and reliability; e) developing alternate forms (A and B); f) providing criterion- and norm-referenced score interpretations; and h) having a set of materials, including manipulatives and links to instructional programs (Let's Think! and Count Too!). Weaknesses of the instrument are related to a) the difficulty in obtaining some of the materials; b) the basis on the Dutch and Finnish curriculums but its use in other countries; and c) limited adaptability, which may result in the need to develop new scales according to the group of children tested.

I Can Do Maths (ICDM)

1. **Needs Addressed**
The purpose of the ICDM is to inform teachers and parents about children's development in numeracy in the early years of schooling. Administering ICDM results in descriptive and normative reports of children's performance in number, measurement, and space (geometry). With these reports, planning a teaching program appropriate to an individual child's needs is made easier. ICDM is a test of beginning mathematical ability in the first three years of school. It is orally administered to students individually or in small groups. It consists of items in the form of pictures. It contains open-ended and multiple-choice questions. The ICDM kit includes two test booklets (level A = pre-K to grade 1; level B = grades 2 and 3), instructions, and norm tables. There are also two equivalent forms. The ICDM items are designed so that the teacher can read the questions aloud while the children use the supplied booklets to mark their responses. For example, the child may be asked to "put a tick under the 10-cent coin" or "put a cross on the shortest snake."

2. **Test Development**
ICDM was originally developed for use with a major national Australian project called *Curriculum and Organization in the Early Years of Schooling*, which investigated the relationships between school entry age, school structures, and later learning outcomes. ICDM was based on theory and extensive literature reviews on how young children learn mathematics. The authors developed test specifications that included several topics (see below).

3. **Competencies Measured**
The ICDM assessment measures the following competencies:

- Number—counting, understand "more" and "less", patterns, 1- to 3-digit numbers, addition, subtraction, order of operation, problem solving

- Measurement—comparing (shortest, smallest, largest), length, calendars, currency (calculate change), clocks, simple graphs

- Space (Geometry)—geometric shapes and solids,

The number of items on each form is 30 in order to a) keep enough items to measure a sample of the content at the three grade levels; and b) reduce administration time.

4. **Administration/Scoring/Reporting**
All items are read to the children to avoid performance being affected by reading factors. Administration time is 20 minutes per student, if given individually. Items may also be administered in a group or class setting. Either of two forms (A and B) may be used since they have been statistically equated. Scoring is conducted in terms of item scores (1 point per item) and raw scores (total scores on all of the items). Scores may then be a) placed on a 0-100 scale; b) reported in terms of percentile ranks for grades 1, 2, and 3; and c) converted to stanines. It is also possible to have subtest scores (number, measurement, and space). There are qualitative descriptions of two score levels, or achievement groups: higher levels of numeracy and lower levels of numeracy. Reliability is high.

5. **Strengths/Weaknesses**
The main strengths of ICDM are a) it tests conceptual and procedural knowledge and skill; b) the manual has information on validity and reliability; c) there are supplemental instructional materials; d) the administration time (20 minutes) is reasonable; and e) it may be administered in a group setting. Weaknesses of the ICDM include a) it is a commercial product requiring ordering and purchasing; b) it has limited content coverage; and c) group administration may have a negative effect on children's performance.

KeyMath Diagnostic Assessment (KeyMath–3)

1. **Needs Addressed**
The first version of KeyMath was developed in Canada in 1971 as a comprehensive mathematics assessment across a broad range of concepts and skills for children in pre-K to grade 6. The latest (fifth) version was developed in 2007 by a) retaining the items that worked well; b) updating the item content; and c) aligning the content with current national mathematics standards. As with the TEAM, it has two levels: preK to grade 2, and grade 3 to grade 6. It assesses learning in three domains: basic concepts (conceptual knowledge), operations (computational skills), and applications (problem solving). It measures mathematical proficiency by providing coverage of the content that is taught in regular mathematics instruction. It was normed in the U.S. and features high content validity (with the National Council of Teachers of Mathematics, or NCTM, standards) and high reliability. Extensive information is available on reliability (test and subtests by form), correlations by subtest, construct validity, and content validity. It is an individually administered assessment that is available in English. KeyMath items are linked with lessons in the KeyMath instructional program.

2. **Test Development**
KeyMath was developed through a review of the NCTM standards. The standards were used to create a test blueprint (specifications) that reflected essential mathematics content and existing curricular priorities. The items went through qualitative (item review) and data based (data review) evaluations. The KeyMath test development process resulted in two test forms (A and B), each with 372 items, which are grouped into 10 subtests (see below for the subtests by domain). Items are identified by the appropriate grade level, so the test takes the following amount of times to administer: pre-K = 15-30 minutes; K = 20-35 minutes; grade 1 = 35-50 minutes; grade 2 = 45-60 minutes; and grade 3 = 60-75 minutes. Testing for the upper grades (4-6) take about 90 minutes.

3. **Competencies Measured**
The KeyMath assessment measures the following competencies:

   - Basic Concepts—numeration, mental computation, estimation
   - Operations—addition, subtraction, multiplication, division, algebra
   - Applications—geometry, measurement, data analysis, probability, problem solving (foundations and applied)

   There are ten subtests: numeration, algebra, geometry, measurement, data analysis/probability, mental computation/estimation, addition/subtraction, multiplication/division, foundations of problem solving, and applied problem solving. According to the authors, the latest version of KeyMath includes more algebraic content to better reflect the latest NCTM standards. More items were added to the lower grades to ensure better measurement at that level. Some items were dropped at the upper grades to reduce testing time without losing measurement precision. A sensitivity review was also conducted to evaluate each item for fairness and appropriateness with respect to gender, ethnicity, and cultural background.

4. **Administration/Scoring/Reporting**
Eight of the subtests are administered with flipbooks and two of the subtests are administered with a written computation examinee booklet. The test is untimed and individually administered. Test administrators need to follow the standardized instructions in the accompanying manual. KeyMath can be scored by hand or by computer entry using the ASSIST scoring and reporting system. ASSIST may be used for a) entering raw scores; b) converting raw scores to derived scores; c) printing and/or saving individual scores and progress reports; d) producing narrative reports that describe a child's results for the test and subtests; e) exporting the scores (for statistical analysis) into another format (text file or Excel spreadsheet). Quantitative analysis can be conducted using raw scores and the following derived scores: scale scores, subtest scale scores, standard scores, percentile ranks, age/grade equivalents, and growth scale values.

5. **Strengths/Weaknesses**
The main strengths of KeyMath are a) large item sets with comprehensive coverage of pre-school and elementary school mathematics; b) linkages to the NCTM standards; c) coverage of concepts and skills; d) ease of administration through flipbooks and instructions; e) ease of scoring and reporting, especially when using the ASSIST system; f) extensive options for reporting; and g) references to the companion instructional system. Weaknesses of the KeyMath include a) relatively long administration time of up to 90 minutes; b) high cost per pupil due to initial outlay and extensive administration time; c) a high number of items that may overwhelm the user.

Test of Early Mathematics Ability (TEMA–3)

1. **Needs Addressed**
The first version of the TEMA (Ginsburg & Baroody, 1983) was developed in response to two needs in early mathematics. These were for a test that would: a) identify children in grades K to 3 with difficulties in learning mathematics; and b) provide useful information about the mathematical strengths and weaknesses of children, both with and without learning difficulties. Since the initial publishing of the TEMA, three additional purposes for the instrument have evolved: a) suggest instructional practices for children; b) document children's progress in learning mathematics; and c) serve as a measure in research projects.

2. **Test Development**
The TEMA went through a specific process in the development of the initial version. Steps in the process included: a) identify items from research studies by the authors and other researchers on children's informal and formal mathematical knowledge; b) pilot the items; c) develop an initial version of the instrument; d) finalize the instrument; e) write instructions for test administration; f) collect reliability and validity data; and g) create a norm group based on a representative sample of children in the U.S. In response to critiques by test reviewers, including in the *Mental Measurements Yearbook*, the instrument has since gone through two major revisions. The current instrument is the TEMA-3.

3. **Competencies Measured**
The latest version of this instrument, the

TEMA-3, has 72 items. This is an increase from the original TEMA, which had 50 items. The items measure children's informal and formal mathematics. Each item is administered individually to the children. Some of the items have multiple parts. The content domains tested by the TEMA-3 are the following (with either informal or formal mathematics in parentheses):

- Numbering (Informal): 23 items
- Number Comparisons (Informal): 6 items
- Calculation (Informal): 7 items
- Concepts (Informal): 4 items
- Numeral Literacy (Formal): 8 items
- Number Facts (Formal): 9 items
- Calculation (Formal): 10 items
- Concepts (Formal): 5 items

There are a total of 40 informal items and 32 formal items. The TEMA documentation gives descriptions of each of the item domains and the individual items. On the test form, the items in the test are arranged in order of difficulty, so that the informal numbering items tend to appear at the beginning of the test while many of the calculation items are more toward the end of the test. The test is not timed.

4. **Administration/Scoring/Reporting**
The TEMA kit consists of an examiner's manual, a picture book, profile/examiner record booklets, and manipulatives (for a few of the items). As with many other oral assessments, the assessor begins with a set of simple instructions. The assessor then fills in information about the child on the instrument and begins with the items. Each item has its own set of instructions. Relative to the other tests reviewed, a unique feature of the TEMA is that there are different entry points depending on the age of the child. The youngest children (age 3) begin with item #1, but children of other ages begin at different points in the test. The assessor also establishes a basal (a lower bound) and a ceiling (an upper bound) for each child. Both the different starting points and the establishment of a basal and ceiling are designed to reduce the administration time while obtaining valid scores. On some of the items, there is a practice item so that the child has a better idea of what they are supposed to do. Each item is scored as either right or wrong. There are criteria for making scoring judgments. These criteria are explained in the examiner's manual, but are also noted on the score sheet. The assessor calculates a total score for each child. The TEMA allows for different types of score interpretations. In addition to a total score, there are tables for calculating grade level equivalents and a math "IQ" score, with a mean of 100 and a standard deviation of 15. These types of norm-referenced scores are perhaps more useful for children in developed countries, but they also provide a yardstick by which to judge the competencies of children in developing countries.

5. **Strengths/Weaknesses**
The main strengths of the TEMA are a) basing the test development in Piagetian and research-based developmental psychology of how young children learn mathematics; b) testing both conceptual and procedural knowledge and skills; c) providing information on both informal and formal mathematics of young children; d) establishing validity and reliability; e) developing alternate forms (A and B); f) providing multiple score interpretations (e.g., scale scores, norm-referenced scores, grade level equivalents, math IQs); g) conducting DIF studies (differential item functioning, to check for item bias); and h) having a comprehensive set of materials, including a booklet on assessment probes and instructional activities. Weaknesses of the instrument are related to its length, including a) a typical administration to a child that can take up to 30 or 40 minutes; b) needing to understand setting a basal and a ceiling, if the child has correctly answered or incorrectly answered 5; c) a lack of a basis in a particular curriculum or set of content standards; d) a problem of non-adaptability, which causes the loss of the scales.

Tools for Early Assessment in Math (TEAM)

1. **Needs Addressed**
TEAM was developed to assess children in all of the five common domains of early mathematics: numbers, operations, geometry, measurement, and patterns. In the documentation, the authors stated that the TEAM cov-

ers areas of mathematics not included in the TEMA-3; they said that the TEMA-3 only covers number competencies but not the other important topics in mathematics (e.g., shapes). TEAM is currently available for children in pre-K to grade 2, and with grades 3 to 5. The TEAM system involves a) assessing students (assess); b) reporting the findings (report); and c) applying the results to instructional practices (apply). It is a one-on-one assessment that is available in English and Spanish.

2. **Test Development**
TEAM went through an extensive development process, including a) the selection of content as valued by mathematicians, educators, and researchers; b) a review of content in the five common domains of early mathematics; c) an examination of the developmental progression of each domain; d) the generation of items by domain and progression; e) pilot testing of 236 items; f) analysis of the items using item response theory; g) elimination of 37 items due to unacceptable item statistics and/or redundancy; h) retention of 199 items; i) selection of items for inclusion on the test form. The items were used in three forms, which take about one hour to administer.

3. **Competencies Measured**
The TEAM assessment measures the following competencies:

   - Numbers—counting, recognition, comparison, sequencing, composition, decomposition
   - Operations—adding, subtracting, multiplying, dividing, fractions
   - Geometry—shape identification, composition, decomposition, comparison, congruence, construction, transformation
   - Measurement—angle, area, length, adding length, units, data
   - Patterns—sequences, missing elements, designs, pre-algebra

   According to the authors, the assessment also covers the "big ideas of mathematics" including a) connecting numbers to the real world; b) proportional reasoning; c) form of a number; d) equality; e) base 10; f) quantity and magnitude; and g) one-to-one correspondence.

4. **.Administration/Scoring/Reporting: Mate**rials for TEAM include flipbooks, manipulatives, scripts, administration guide, and online access for scoring, tracking, and reporting. All items are scored either correct or incorrect (one point each). Answers may be recorded on a computer, hand-held device, or score sheet. Reports include a) individual student results; b) interpretations according to performance standards; and c) progress summaries. There are suggested lesson plans that map to the curriculum and student results. Reports can be generated for students, classrooms, schools, and districts. Reports can also include grade level equivalents, percent correct scores, performance categories, performance level descriptors, and student trajectories. Specific lessons and activities are provided for the users so that the children's weaknesses may be addressed.

5. **Strengths/Weaknesses**
The main strengths of the TEAM assessment are a) comprehensive coverage of early math domains; b) linkages to content standards; c) coverage of concepts and procedural knowledge; d) ease of administration through flipbooks and instructions; e) ease of scoring and reporting, especially when using technology; and f) extensive research and pilot testing. Weaknesses of the TEAM include a) relatively long administration time of one hour; b) high cost per pupil due to extensive administration time; c) lack of documentation available on the web (i.e., specific instructions need to be purchased); d) setting a basal and ceiling could be confusing to the assessors; e) questionable value-added beyond the TEMA-3.

Trends in International Mathematics and Science Study (TIMSS)

1. **Needs Addressed**
For the past 20 years, the TIMSS has measured trends in mathematics (and science) achievement at the fourth and eighth grades of over 60 developed and developing countries. It has been conducted on a regular four-year cycle during that time period. Countries pay to participate in the TIMSS, both through fees to the IEA as well as in local operational costs. Countries use it in various ways: a) monitoring system-level achievement trends in a global context; b) establishing achievement

goals and standards for educational improvement; c) stimulating curriculum reform; d) improving teaching and learning through research and analysis of data; e) conducting related studies, such as monitoring equity; and f) training researchers and teachers in assessment and evaluation. Over 600,000 students participated in the 2011 TIMSS. The next TIMSS administration, in 2015, will include TIMSS Numeracy (described in the next section).

2. **Test Development**
TIMSS sends curriculum questionnaires to participating countries in order to gather information about the mathematics subject matter that is a) intended from a national perspective b) implemented in the schools and classrooms; and c) attained by the students of different characteristics. Based on this information, TIMSS curriculum experts develop assessment frameworks. These frameworks specify the types of items that will be developed for the tests. About half of the items developed and used on the tests are multiple-choice, and the other half are constructed response items where the children write their answers. A total of 175 mathematics items were used for grade 4 in 2011.

3. **Competencies Measured**
The TIMSS grade 4 content domains are the following:

   - Number—whole numbers; fractions and decimals; number sentences with whole numbers; and patterns and relationships (50%)
   - Geometric Shapes and Measures—points, lines, and angles; and two- and three-dimensional shapes (35%)
   - Data Display—reading and interpreting data; and organizing and representing data (15%)

   The TIMSS items are also classified by cognitive domains, which describe the type of thinking required to answer the items:

   - Knowing (40%)
   - Applying (40%)
   - Reasoning (20%)

   TIMSS uses a matrix sampling system in which different items are administered to different students, with the goal of creating scores for the country rather than for individual students. In addition to the test items, TIMSS administers extensive surveys to create contextual data. Survey categories are a) early exposure (pre-school and home); b) home resources; c) school resource levels; d) school environments; e) teacher preparation; f) teacher career satisfaction; g) student attitudes; h) engaging instruction; and i) student nutrition and sleep.

4. **Administration/Scoring/Reporting**
There are 14 different TIMSS mathematics booklets. Each student takes 25 items out of the 175 total items, so that each item appears in two booklets. Assessment time for grade 4 students is approximately 72 minutes. An additional 30 minutes is allocated for administering the student questionnaires. Each multiple-choice item is scored as one point, and each constructed response item is scored as either one or two points depending on the nature of the item. TIMSS uses item response theory (IRT) to analyze the test results, which are reported on the TIMSS achievement scale, with a range of 0 to 1,000 (student performance generally ranges from 300 to 700). TIMSS uses the center-point of the scale (500) as a point of reference that remains constant from assessment to assessment. Scores are reported in terms of a) averages (means) and score distributions; b) trends across time (relative to previous administrations in 1995, 1999, 2003, and 2007); c) trends across grade levels (fourth to eighth); d) achievement by gender; and e) performance benchmarks (advanced, high, intermediate, and low). TIMSS results are disseminated through reports and via the web through a well-documented international database for within and across country research and evaluation. In addition, some items are released into the public domain so that researchers and others may use them.

5. **Strengths/Weaknesses**
The main strengths of TIMSS are a) high quality assessments based on an internationally-derived curriculum; b) participation by over 60 countries; c) high level of expertise in scoring and reporting; d) contextual analysis based on questionnaire data; e) testing both concepts and skills; f) establishing strong validity and reliability; f) providing score

reporting with trends over time. Weaknesses are related to a) test administration every four years; b) high cost for participating countries; c) comparisons that sometimes reflect poorly on the country; d) lack of full test release.

TIMSS Numeracy

1. **Needs Addressed**
   A new assessment from IEA, TIMSS Numeracy will be conducted on a regular four-year cycle starting in 2015. As with the regular TIMSS, countries will pay to participate in the TIMSS, both through fees to the IEA as well as in local operational costs. Countries will use it in the same ways: a) monitoring system-level achievement trends in a global context; b) establishing achievement goals and standards for educational improvement; c) stimulating curriculum reform; d) improving teaching and learning through research and analysis of data; e) conducting related studies, such as monitoring equity; and f) training researchers and teachers in assessment and evaluation. The differences between TIMSS and TIMSS Numeracy are a) lower level of knowledge and skills required; b) oriented more towards developing countries where students have difficulty with the TIMSS curriculum and tests; and c) flexible grade levels so that it can be administered to students at grades 4, 5, or 6. There is an equivalent for reading, called prePIRLS, which will be administered starting in 2016 (in conjunction with the 5-year cycles for PIRLS).

2. **Test Development**
   As with TIMSS, the TIMSS Numeracy developers send curriculum questionnaires to participating countries in order to gather information about the mathematics subject matter that is a) intended from a national perspective b) implemented in the schools and classrooms; and c) attained by the students of different characteristics. Based on this information, TIMSS Numeracy experts develop assessment frameworks that specify the types of items that will be developed for the tests. About half of the items developed and used on the tests will be multiple-choice, and the other half are constructed response items where the children write their answers.

3. **Competencies Measured**
   The grade 4 TIMSS Numeracy content domains are the following:

   - Whole Numbers—place value; recognizing and writing; comparing and ordering; computing; solving word problems
   - Fractions—recognizing simple fractions; representing fractions using words, numbers, or models
   - Geometric Shapes and Measures—points, lines, and angles; two- and three-dimensional shapes; measuring and estimating length
   - Data Display—reading data from tables, bar graphs, and pictographs; solving simple problems The TIMSS items are also classified by cognitive domains, which describe the type of thinking required to answer the items:

   As with TIMSS, the TIMSS Numeracy assessments will use a matrix sampling system in which different items are administered to different students, with the goal of creating scores for the country rather than for individual students. In addition to the test items, TIMSS Numeracy will administer surveys to create contextual data in categories such as a) early exposure (pre-school and home); b) home resources; c) school resource levels; d) school environments; e) teacher preparation; f) teacher career satisfaction; g) student attitudes; h) engaging instruction; and i) student nutrition and sleep.

4. **Administration/Scoring/Reporting**
   There will be different TIMSS Numeracy booklets, and each student will take a subset of the total items. Multiple choice items will be scored as one point each, and constructed response items will be scored as either one or two points depending on the nature of the item. TIMSS Numeracy will use item response theory (IRT) to analyze the test results and report them on the TIMSS achievement scale, with a range of 0 to 1,000 (student performance generally ranges from 300 to 700). TIMSS Numeracy will use the center-point of the scale (500) as a point of reference that remains constant from assessment to assessment. Similarly to TIMSS, the TIMSS Numeracy scores will be reported in terms of a) averages (means) and score distributions; b) achievement by gender; and c) performance benchmarks (advanced, high, intermediate, and low). As TIMSS Numeracy is continued after 2015, trends will be reported in the fu-

ture. The results will be disseminated through reports and via the web through a well-documented international database for within and across country research and evaluation. In addition, some items will be released into the public domain so that researchers and others may use them.

5. **Strengths/Weaknesses**
The main strengths of TIMSS Numeracy will be the same as those for TIMSS: a) high quality assessments based on an internationally-derived curriculum; b) participation by multiple countries; c) high level of expertise in scoring and reporting; d) contextual analysis based on questionnaire data; e) testing both concepts and skills; and f) establishing strong validity and reliability. Weaknesses are related to a) test administration every four years; b) high cost for participating countries; c) comparisons that sometimes reflect poorly on the country; d) lack of full test release. IEA is providing financial incentives to countries who would like to participate in the first TIMSS Numeracy assessment in 2015.

Gaps in the Existing Landscape

There are two main gaps in the tests covered in this paper, both of which could be corrected at a later time. First, the survey was limited to tests published in English. There are surely dozens of other tests available in other languages. Second, there is no information in this paper on formative assessment, which is perhaps the most promising assessment avenue for improving children's learning. Formative assessment was covered extensively in the previous paper (GIZ, 2012).

Otherwise, the ten tests reviewed for this paper offer a range of alternatives, with tools of shorter length (ASER, Uwezo, ICDM), medium length (EGMA, TEMA, ENT, TEAM), and longer length (Keymath, TIMSS, TIMSS Numeracy). The length of the test generally corresponds to the content coverage, and somewhat to the grade level. All of the tests should be adaptable to other languages and cultures, and many of them have already been translated and/or used in developing countries (ASER, Uwezo, EGMA, ENT, TEMA, TIMSS, and TIMSS Numeracy).

Other numeracy assessments for future study, though again somewhat restricted to English-language tests, may include the following:

1. Performance Indicators in Primary School (PIPS) On-Entry Baseline Assessment
2. New Zealand School Entry Assessment (SEA) – Numeracy
3. Early Math Diagnostic Assessment (EMDA)
4. Process Assessment of the Learner II – Mathematics (PAL-II Math)
5. Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Math

Summary and Recommendations

Summary

The two objectives of the study – developing a **format** for presenting information on assessment tools and provide a **review** of the landscape of numeracy assessment tools – were achieved. The descriptions and criteria provide a system for identifying and classifying the characteristics of the assessments. The review of the landscape was completed, under the limitations of the study, though many instruments, especially those in languages other than English, were not addressed.

One way to summarize the findings is through the learning domains proposed by the Learning Metrics Task Force (LMTF) that is coordinated by the UNESCO Institute of Statistics (UIS) and the Brookings Institution Center for Universal Education (CUE). In their system of content standards, there are six domains of mathematics for pre-school and primary school children. The domains covered by the ten assessments in this report are provided in the table below.

It is clear that not all assessments have items that relate to the LMTF domains. However, it is not expected that this would be the case. Many of the assessments fulfill a different purpose, whether through shorter length or a focus on the numeracy part of mathematics. The more comprehensive (and longer) assessments – such as KeyMath, TEAM, TIMSS, and TIMSS Numeracy – are more comprehensive in their coverage.

While this is the current status of the assessments, some (e.g., EGMA) have provisions for covering other areas that are not in their core instrument (e.g., spatial and relational reasoning). The table (and descriptions) will need to be updated on a regular basis.

Tools	LMTF Pre-Primary and Primary School Domains					
	Number Sense	Operations	Spatial Sense and Geometry	Patterns and Classification	Measurement and Comparison	Math Applications
ASER	X	X				
Uwezo	X	X				
EGMA	X	X		X		X
ENT	X			X	X	
ICDM	X	X	X	X	X	X
KeyMath	X	X	X		X	X
TEMA	X	X		X		X
TEAM	X	X	X	X	X	X
TIMSS	X	X	X	X		X
TIMSS Numeracy	X	X	X	X	X	X

Recommendations

It is clear that more work needs to be done with the mathematics assessment tools and information. Some of these areas are the following:

1. The criteria need to be reviewed by mathematics experts.

2. The assessment information needs to checked and completed.

3. Other assessments should be added to the database.

4. The usefulness of the information needs to be evaluated.

5. Linkages with other systems (e.g., LMTF and OLO) need to be improved.

The goal is for the practitioners and experts to use the data to find ways of selecting instruments that can help children to improve their learning. The hope is that the present study has contributed to this effort.

References

AERA, APA, & NCME (1999)
The Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Aunio, P. (2006)
Number Sense in Young Children – International Group Differences and an Intervention Programme for Children with Low and Average Performance (Research Report 269). Helsinki, Finland: University of Helsinki.

Aunio, P. et al. (2009)
Early Numeracy in Low-Performing Young Children. British Educational Research Journal, 35:1, 25-46.

Aunio, P. et al. (2008)
Children's Early Numeracy in England, Finland, and People's Republic of China. *International Journal of Early Years Education*, 16:3, 202-221.

Clements, D. & Sarama, J. (2008)
Development of a Measure of Early Mathematics Achievement Using the Rasch Model: The Research-Based Early Math Assessment. *Educational Psychology*, 28:4, 457-482.

Connolly, A. (2007)
KeyMath–3 Diagnostic Assessment. Richmond Hill, Ontario: Psycan.

Doig, B. & De Lemos, M. (2000)
I Can Do Maths Teacher's Guide. Camberwell, Victoria, Australia: ACER Press.

Ginsburg, H. P., & Baroody, A. J. (1983)
The Test of Early Mathematics Ability. Austin, TX: Pro-Ed.

Ginsburg, H. P., & Baroody, A. J. (2003)
The Test of Early Mathematics Ability (3rd ed.). Austin, TX: Pro-Ed.

GIZ/BMZ (2012)
Learning Outcomes Assessments and Numeracy with Reference to Early Grade Numeracy in Low Income Countries. Bonn/Eschborn.

International Test Commission (2000)
International Guidelines for Test Use.
ITC: *www.intestcom.org/itc_projects.htm*

Mullis, I.V.S., Martin, M.O., Foy, P., and Arora, A. (2012)
TIMSS 2011 *International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS Study Center, Boston College.

National Governors Association Center for Best Practices (2010)
Common Core State Standards for Mathematics. Washington, DC: NGACBP.

National Council of Teachers of Mathematics (2000)
Principles and Standards for School Mathematics. Reston, Virginia: NCTM.

Reubens, A. (2009)
Early Grade Mathematics Assessment (EGMA): A Conceptual Framework Based on Mathematics Skills Development in Children. Research Triangle Park, NC: EdData II.

Annex: Evaluation Criteria

Annual Status of Education Report – Mathematics (ASER)

1. Overview
    a. Type: diagnostic
    b. Target group: ages 5 to 16
    c. Administration: individual (oral)
    d. Developer: Pratham
    e. Accessibility: open
    f. Cost structure: centralized test development and reporting; village volunteers administer and score tests
    g. Countries: India, Pakistan
2. Content
    a. Learning domains: number recognition, operations, others
    b. Cognitive domains: knowledge, comprehension, application
    c. Item type: objective
    d. Number of items: several tasks with multiple items per task
    e. Adaptation to local context: minimal
    f. Alternate forms: no
3. Implementation
    a. Materials: administration guide, scoring sheets
    b. Administration time: 10 minutes per student
    c. Enumerator training: administration guide, video
    d. Timed items: no
    e. Technology-based application: no
4. Analysis and reporting
    a. Scoring: raw scores
    b. Reliability: not reported
    c. Score reports: village report cards (by the village volunteers)
    d. Scaling: percent correct scores
    e. Performance categories: yes
    f. Performance level descriptors: no
    g. Computer application: no
5. Recommendations
    a. Strengths: easy to administer, minimal training required, immediate feedback, simple tasks, low cost, readily scalable, involvement by parents
    b. Weaknesses: focus on procedural knowledge, little emphasis on concepts, no informal math, possible problems with standardization of administration
    c. Suggestions: modify tools to include addition and multiplication; increase standardization; expand reports to include psychometrics (item statistics, reliability, performance level descriptors)

6. Further information
    a. Assistance: Pratham/ASER Centre, SAFED (South Asian Forum for Educational Development)/ASER Pakistan
    b. Point of contact: ASER Centre, B 4/54 Safdarjung Enclave, New Delhi, India 110 029, +91 11 4602 3612 / 2671 6084; ASER Pakistan, 41 – L Model Town Ext., Lahore, Pakistan, +92 (42) 351 73005-7
    c. Link (website): *www.pratham.org*; *www.asercentre.org*; *www.aserpakistan.org*
7. User feedback
    a. Rating system (five stars):
    b. Comments:

"Capability" in Kiswahili (Uwezo) - Mathematics

1. Overview
    a. Type: diagnostic
    b. Target group: ages 5 to 16
    c. Administration: individual (oral)
    d. Developer: UWEZO (based on Pratham's work on ASER)
    e. Accessibility: open
    f. Cost structure: centralized test development and reporting; village volunteers administer and score tests
    g. Countries: Kenya, Tanzania, Uganda (also Mali, Senegal)
2. Content
    a. Learning domains: number recognition, operations, others
    b. Cognitive domains: knowledge, comprehension, application
    c. Item type: objective
    d. Number of items: several tasks with multiple items per task
    e. Adaptation to local context: some differences across countries
    f. Alternate forms: no
3. Implementation
    a. Materials: administration guide
    b. Administration time: 10 to 15 minutes per student
    c. Enumerator training: administration guide, video
    d. Timed items: no
    e. Technology-based application: no
4. Analysis and reporting
    a. Scoring: raw scores
    b. Reliability: not reported
    c. Score reports: village report cards (by the village volunteers)
    d. Scaling: percent correct scores
    e. Performance categories: yes
    f. Performance level descriptors: no
    g. Computer application: no
5. Recommendations
    a. Strengths: easy to administer (by volunteers), minimal training required, immediate feedback, low cost, easily scalable, parental involvement
    b. Weaknesses: focus on procedural knowledge, little emphasis on concepts, no informal math, possible problems with standardization of administration
    c. Suggestions: modify tools to include addition and multiplication; increase standardization; expand reports to include psychometrics (item statistics, reliability, performance level descriptors)

6. Further information
    a. Assistance: UWEZO (see below)
    b. Point of contact: UWEZO East Africa, 3rd floor, ACS Plaza, Lenana Road, PO Box 19875 – 00200, Nairobi, Kenya, +254 20 386 1372/3/4
    c. Link (website): *www.uwezo.net*
7. User feedback
    a. Rating system (five stars):
    b. Comments:

Early Grade Math Assessment (EGMA)

1. Overview
    a. Type: diagnostic
    b. Target group: grades 1-3
    c. Administration: individual (oral)
    d. Developer: RTI International, USAID, World Bank
    e. Accessibility: open
    f. Cost structure: administration costs
    g. Countries: More than ten developing countries (see tracker at *www.eddataglobal.org*)
2. Content
    a. Learning domains: counting, number sense, operations, problem solving
    b. Cognitive domains: knowledge, comprehension, application
    c. Item type: objective
    d. Number of items: 8 tasks (maximum) with multiple items per task
    e. Adaptation to local context: yes
    f. Alternate forms: depends on situation
3. Implementation
    a. Materials: instruments, reports, guidance notes, webinar
    b. Administration time: 15 minutes per student
    c. Enumerator training: guidance notes, webinar
    d. Timed items: yes and no
    e. Technology-based application: tablet, laptop (Tangerine, eEGMA)
4. Analysis and reporting
    a. Scoring: raw scores
    b. Reliability: sometimes reported (coefficient alpha, not test-retest)
    c. Score reports: nothing programmed
    d. Scaling: percent correct scores possible
    e. Performance categories: possible but not yet developed
    f. Performance level descriptors: no
    g. Computer application: not fully developed; beta version of application to help in determining cut scores
5. Recommendations
    a. Strengths: high face validity, universal tasks, possibility for context-specific adaptation, measurement of concepts and skills, rapid assessment
    b. Weaknesses: high per pupil costs, lack of alignment to the curriculum, no informal math, extensive training required for testing, weak sustainability
    c. Suggestions: figure out how to get more countries to take the lead, focus more on use of information for instructional improvement
6. Further information
    a. Assistance: RTI International. EdData II (see below)

b. Point of contact: RTI International, EdData II, 3040 Cornwallis Road, PO Box 12194, Research Triangle Park, North Carolina 27709-2194 USA, +1 919-541-6000

c. Link (website): www.rti.org; www.eddataglobal.org

7. User feedback

a. Rating system (five stars):

b. Comments:

Early Numeracy Test (ENT)

1. Overview

a. Type: diagnostic

b. Target group: ages 3 to 8 years

c. Administration: individual (oral)

d. Developer: originally created as the Dutch Utrecht's Early Numeracy Test (Van Luit and Vand De Rijt, 1994) and then adapted other contexts and languages, including Finnish (Aunio, 2006)

e. Accessibility: open

f. Cost structure: administration costs

g. Countries: Belgium, China, Finland, Germany, Greece, Hong Kong, Netherlands, Singapore, Slovenia, Spain, U.K.

2. Content

a. Learning domains: mathematical prerequisites (comparison, classification, one-to-one correspondence, and number series) and counting skills (number words, counting, number sense)

b. Cognitive domains: knowledge, comprehension, application

c. Item type: objective

d. Number of items: 40 items; all items scored as correct or incorrect

e. Adaptation to local context: slight adaptation possible (but generally not much adaptation is needed)

f. Alternate forms: 3 forms (A, B, and C – form C is made up of items from forms A and B)

3. Implementation

a. Materials: administration guide, paper and pencil, pictures, record sheets, manipulatives (cubes)

b. Administration time: 30 minutes per student

c. Enumerator training: administration guide

d. Timed items: no

e. Technology-based application: no

4. Analysis and reporting

a. Scoring: raw scores, item scores

b. Reliability: yes (coefficient alpha)

c. Score reports: no

d. Scaling: norms (Finnish), percent correct scores

e. Performance categories: no

f. Performance level descriptors: no

g. Computer application: no

5. Recommendations

a. Strengths: research-based (from different countries), tests conceptual and procedural knowledge and skill, comprehensive information on validity and reliability, international comparisons, different languages, supplemental instructional materials (Let's Think! and Count Too!)

b. Weaknesses: high cost per pupil due to administration time, not curriculum based, not adaptable (loss of scaling)

c. Suggestions: test has been modified over time and has been made international; some limitations for the older children (ceiling effects reported)

6. Further information

a. Assistance: None

b. Point of contact: Pirjo Aunio, Niilo Maki Institute, PO Box 35, 40014, University of Jyvaskyla, Finland, +35 850-43-43-408

c. Link (email): *pirjo.aunio@mnmi.fi*

7. User feedback

a. Rating system (five stars):

b. Comments:

I Can Do Maths (ICDM)

1. Overview

a. Type: diagnostic

b. Target group: first three years of school

c. Administration: individual or small group

d. Developer: Brian Doig and Marion De Lemos (ACER Press, 2000)

e. Accessibility: open (for purchase)

f. Cost structure: commercial ($61.95 for kit) plus administration costs

g. Countries: Australia

2. Content

a. Learning domains: number, space, measurement

b. Cognitive domains: knowledge, comprehension, application

c. Item type: multiple-choice and open-ended

d. Number of items: 30 items

e. Adaptation to local context: no

f. Alternate forms: 2 forms (A and B) Implementation

g. Materials: manual, test booklets (A and B), specimen kit

h. Administration time: 20 minutes per student

i. Enumerator training: administration guide

j. Timed items: no

k. Technology-based application: no

3. Analysis and reporting

a. Scoring: raw scores, item scores

b. Reliability: yes (coefficient alpha)

c. Score reports: yes

d. Scaling: scale scores (0-100), stanines

e. Performance categories: no

f. Performance level descriptors: qualitative descriptions based on scores

g. Computer application: no

4. Recommendations

a. Strengths: tests conceptual and procedural knowledge and skill, information on validity and reliability, supplemental instructional materials, reasonable administration time (and may be administered in small groups), low cost

b. Weaknesses: commercial product requiring initial purchase, limited content coverage, not curriculum based, not adaptable (loss of scaling)

c. Suggestions: offer more description of test

5. Further information

a. Assistance: ACER (Australian Council for Educational Research)

b. Point of contact: ACER (publisher), 19 Prospect Hill Rd, Camberwell, VIC 3124 Australia, +61 3 9277 5447

c. Link (website): www.shop.acer.edu.au

6. User feedback

a. Rating system (five stars):

b. Comments:

KeyMath Diagnostic Assessment (KeyMath–3)

1. Overview

a. Type: diagnostic

b. Target group: grades K to 12

c. Administration: individual (oral); 8 subtests administered with a flip easel and 2 subtests administered with a written booklet

d. Developer: Austin Connolly (3rd edition in 2007; 1st edition was in 1971)

e. Accessibility: open (for purchase)

f. Cost structure: commercial ($789.00 for kit) plus administration costs

g. Countries: U.S.

2. Content

a. Learning domains: basic concepts (numeration, algebra, geometry, measurement, data analysis and probability0; operations (mental computation and estimation, addition, subtraction, multiplication, division); applications (foundations of problem solving, applied problem solving)

b. Cognitive domains: knowledge, comprehension, application

c. Item type: objective

d. Number of items: 372 items per form grouped into 10 subtests

e. Adaptation to local context: slight adaptation possible (not much needed)

f. Alternate forms: 2 forms (A and B)

3. Implementation

a. Materials: examiner's manual, flip easel, written booklet, record booklets

b. Administration time: 30 to 75 minutes per student (ranging from pre-K at 15-30 minutes to grade 3 at 60-75 minutes)

c. Enumerator training: administration guide, webinar

d. Timed items: no

e. Technology-based application: no

4. Analysis and reporting

a. Scoring: raw scores, item scores

b. Reliability: yes (coefficient alpha)

c. Score reports: yes

d. Scaling: grade- and age-level equivalents, percentile ranks, growth scale values (all norm-referenced), percent correct scores for test and subtests

e. Performance categories: yes

f. Performance level descriptors: yes (including narrative descriptions)

g. Computer application: yes (ASSIST scoring software) or manual scoring

5. Recommendations

a. Strengths: based on National Council of Teachers of Mathematics (NCTM) standards, linked to the KeyMath instructional program, comprehensive tool that measures concepts and skills taught in math instruction, requires fairly extensive training to administer, has parallel forms

b. Weaknesses: high cost per pupil due to administration time, commercial product requiring initial purchase, not adaptable (loss of scaling)

c. Suggestions: toolkit has been modified over time to correct issues

6. Further information
    a. Assistance: Pearson
    b. Point of contact: Pearson, 19500 Bulverde Road, PO Box 599700, San Antonio, Texas 78259, +1 800-627-7271, 800-622-3231
    c. Link (website): www.pearsonclinical.com/education/products
7. User feedback
    a. Rating system (five stars):
    b. Comments:

Test of Early Mathematics Ability (TEMA–3)

1. Overview
    a. Type: diagnostic
    b. Target group: ages 3 to 8
    c. Administration: individual (oral)
    d. Developer: Herbert Ginsburg and Arthur Baroody (Pro-Ed; 3rd edition in 2004; 1st edition in 1983)
    e. Accessibility: open (for purchase)
    f. Cost structure: commercial ($321.00 for kit) plus administration costs
    g. Countries: Benin, Haiti, Macedonia, South Korea, U.S.
2. Content
    a. Learning domains: numbering, numerals, number comparisons, number facts, calculation, concepts (informal and formal mathematics)
    b. Cognitive domains: knowledge, comprehension, application
    c. Item type: objective
    d. Number of items: 72 items (40 informal and 32 formal), some items have multiple parts; all items scored as correct or incorrect
    e. Adaptation to local context: slight adaptation possible (but generally not much adaptation is needed)
    f. Alternate forms: 2 forms (A and B)
3. Implementation
    a. Materials: examiner's manual, picture books (A and B), record booklets, manipulatives, assessment probes and instructional activities booklet
    b. Administration time: 30 minutes per student
    c. Enumerator training: examiner's manual
    d. Timed items: no
    e. Technology-based application: no
4. Analysis and reporting
    a. Scoring: raw scores, item scores
    b. Reliability: yes (coefficient alpha)
    c. Score reports: no
    d. Scaling: grade level equivalents, age equivalents, math "IQ" scores (all norm-referenced), percent correct scores
    e. Performance categories: no
    f. Performance level descriptors: no
    g. Computer application: no
5. Recommendations
    a. Strengths: research-based (Piagetian and other), tests conceptual and procedural knowledge and skill, information on informal and formal math, comprehensive information on validity and reliability, supplemental assessment probes and instructional materials
    b. Weaknesses: high cost per pupil due to administration time, commercial product requiring initial purchase, setting basal and ceiling can be confusing to enumerators, not curriculum based, not adaptable (loss of scaling)

c. Suggestions: toolkit has been modified over time to correct issues

6. Further information

a. Assistance: Pro-Ed (see below)

b. Point of contact: Pro-Ed (publisher), 8700 Shoal Creek Boulevard, Austin, Texas 78757-6897, +1 512 451 3246

c. Link (website): *www.proedinc.com*

7. User feedback

a. Rating system (five stars):

b. Comments:

Tools for Early Assessment in Math (TEAM)

1. Overview

a. Type: diagnostic

b. Target group: pre-K to grade 2

c. Administration: individual (oral)

d. Developer: Douglas Clements, Julie Sarama (McGraw-Hill)

e. Accessibility: open (if purchased)

f. Cost structure: commercial ($289.80 for kit); online license $2 per student

g. Countries: U.S.

2. Content

a. Learning domains: numbers (counting, comparing, recognizing, and subitizing); operations (addition, subtracting, multiplying, dividing); fractions, classifying and analyzing data; measurement (angle, area, length); shapes (comparing, recognizing, composing); spatial sense; patterns and pre-algebraic thinking

b. Cognitive domains: knowledge, comprehension, application

c. Item type: objective

d. Number of items: 199 items (selected out of 236 items originally proposed)

e. Adaptation to local context: slight adaptation possible (but generally not much adaptation is needed)

f. Alternate forms: 2 forms (A and B)

3. Implementation

a. Materials: teacher's guide, flip book (with teacher script), scoring sheets, manipulatives (cubes, chips, coins, etc.), sampler

b. Administration time: 30-40 minutes per student

c. Enumerator training: administration guide, webinar

d. Timed items: no

e. Technology-based application: yes

4. Analysis and reporting

a. Scoring: raw scores, item scores (basal and ceiling)

b. Reliability: yes (coefficient alpha)

c. Score reports: yes (individual, class, district); coding with content standards; links to lessons

d. Scaling: grade level equivalents, percent correct scores

e. Performance categories: yes (above level, at level, below level)

f. Performance level descriptors: yes

g. Computer application: handheld, laptop (including online access for scoring, tracking, and reporting)

5. Recommendations

a. Strengths: wide range of math items, numeracy and other math topics, research-based, linked to content standards, tests conceptual and procedural knowledge, comprehensive information on validity and reliability, supplemental instructional materials (assess, report, apply)

   b. Weaknesses: high cost per pupil due to administration time, commercial product requiring initial purchase, setting basal and ceiling can be confusing to enumerators, not adaptable (loss of scaling)

   c. Suggestions: significant work has taken place to select, eliminate, and modify items

6. Further information

   a. Assistance: McGraw-Hill Customer Service

   b. Point of contact: McGraw-Hill School Education, PO Box 182605, Columbus, Ohio 43218, +1 877-833-5524/800-334-7344

   c. Link (website): *www.mheonline.com/program/view/4/7/335/007TEAM*; *www.orders_mhe@mheducation.com*

7. User feedback

   a. Rating system (five stars):

   b. Comments:

Trends in International Mathematics and Science Study (TIMSS)

1. Overview

   a. Type: international summative (every 4 years; next testing in 2015)

   b. Target group: grade 4

   c. Administration: group (paper and pencil)

   d. Developer: International Association for the Evaluation of Educational Achievement (IEA)

   e. Accessibility: semi-open (secure and released items)

   f. Cost structure: fee-based ($25,000 and 25,000 Euro plus local costs), based on number of students and local costs

   g. Countries: 58 countries currently signed up for 2015

2. Content

   a. Learning domains: numbers, operations, pre-algebra, geometry,

   b. Cognitive domains: knowing, applying, reasoning

   c. Item type: objective (multiple choice and short answer)

   d. Number of items: matrix items, so that one student receives about 40 items (200 items total)

   e. Adaptation to local context: adaptation from country to country with maintenance of content for each item

   f. Alternate forms: various forms with matrix sampling

3. Implementation

   a. Materials: examiner's manual, student booklets

   b. Administration time: 60 minutes per student

   c. Enumerator training: administration guide, video, webinar

   d. Timed items: no

   e. Technology-based application: no

4. Analysis and reporting

   a. Scoring: raw scores, item scores

   b. Reliability: yes (coefficient alpha)

   c. Score reports: yes (country-level)

   d. Scaling: raw scores, scale scores

   e. Performance categories: yes (advanced, high, medium, and low)

   f. Performance level descriptors: yes

   g. Computer application: yes (reports generated by IEA; some training for country partners)

5. Recommendations

    a. Strengths: valuable vehicle for studying international trends in mathematics, high quality, tests conceptual and procedural knowledge and skill, comprehensive information on validity and reliability, extensive technical documentation, curriculum analysis, comparisons with other countries

    b. Weaknesses: high cost per pupil due to processing fees, though a lot is provided for the cost, overemphasis on cross-country comparisons

    c. Suggestions: countries should focus less on international comparisons (which draw the most attention) and concentrate on successes and improvement within the country (or by states or provinces within the country)

6. Further information

    a. Assistance: IEA (see below), Boston College (Massachusetts, U.S.), Data Processing Center (Hamburg, Germany)

    b. Point of contact: IEA, Herengracht 487, 1017 BT Amsterdam, The Netherlands, +31 20 625 3625

    c. Link (website): *www.iea.nl*

7. User feedback

    a. Rating system (five stars):

    b. Comments:

TIMSS Numeracy

1. Overview

    a. Type: international summative (every 4 years; initial testing in 2015)

    b. Target group: grade 4, 5, or 6

    c. Administration: group (paper and pencil)

    d. Developer: International Association for the Evaluation of Educational Achievement (IEA)

    e. Accessibility: semi-open (secure and released items)

    f. Cost structure: fee-based ($25,000 and 25,000 Euro plus local costs), based on number of students and local costs

    g. Countries: to be determined

2. Content

    a. Learning domains: whole numbers, fractions, geometric shapes and measures, and data display

    b. Cognitive domains: knowledge, comprehension, application

    c. Item type: objective (multiple choice and short answer)

    d. Number of items: matrix items, so that one student receives about 40 items

    e. Adaptation to local context: adaptation from country to country with maintenance of content for each item

    f. Alternate forms: various forms with matrix sampling

3. Implementation

    a. Materials: examiner's manual, student booklets

    b. Administration time: 60 minutes per student

    c. Enumerator training: administration guide

    d. Timed items: no

    e. Technology-based application: no

4. Analysis and reporting

    a. Scoring: raw scores, item scores

    b. Reliability: yes (coefficient alpha)

    c. Score reports: yes (country-level)

    d. Scaling: raw scores, scale scores

    e. Performance categories: yes (advanced, high, medium, and low)

f. Performance level descriptors: yes

g. Computer application: no

5. Recommendations

a. Strengths: assesses fundamental knowledge, procedures, and problem-solving strategies, valuable vehicle for studying international trends in mathematics, high quality, comprehensive information on validity and reliability, extensive technical documentation, presentations, country-level curriculum and instructional analysis, comparisons with other countries

b. Weaknesses: high cost per pupil due to processing fees, though probably a reasonable value for funds spent

c. Suggestions: countries should focus less on international comparisons (which draw the most attention) and concentrate on successes and improvement

6. Further information

a. Assistance: Assistance: IEA (see below), Boston College (Massachusetts, U.S.), Data Processing Center (Hamburg, Germany)

b. Point of contact: IEA, Herengracht 487, 1017 BT Amsterdam, The Netherlands, +31 20 625 3625

c. Link (website): *www.iea.nl*

7. User feedback

a. Rating system (five stars):

b. Comments:

Published by
Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Registered offices
Bonn and Eschborn,
Germany

Sector Programme Numeracy
Education Section
Division of Education, Health and Social Protection
Godesberger Allee 119
53175 Bonn
Germany
Tel. +49 (0) 228 44 60 - 0
Fax +49 (0) 228 44 60 - 1766

Dag-Hammarskjöld-Weg 1-5
65760 Eschborn
Germany
Tel. +49 (0) 6196 79 - 0
Fax +49 (0) 6196 79 - 1115

numeracy@giz.de
www.giz.de/numeracy

Author
Jeff Davis (Education Development Associates LLC)

Design and Layout
Diamond media GmbH, Neunkirchen-Seelscheid

Printed by
druckriegel GmbH, Frankfurt/Main, Germany
Printed on FSC-certified paper

Photo credits
© GPE / Deepa Srikantaiah

As at
November 2014

On behalf of
Federal Ministry for Economic Cooperation and Development (BMZ)
Division of Education and the Digital World

Adresses of the BMZ offices
BMZ Bonn
Dahlmannstraße 4
53113 Bonn
Germany
Tel. +49 (0) 228 99 535 - 0
Fax +49 (0) 228 99 535 - 3500

BMZ Berlin
Stresemannstraße 94
10963 Berlin
Germany
Tel. +49 (0) 30 18 535 - 0
Fax +49 (0) 30 18 535 - 2501

poststelle@bmz.bund.de
www.bmz.de