Using EGRA to Report on SDG 4.1.1a

10th Meeting of GAML 2023 - Paris, France
Melissa Chiappetta
USAID
AGENDA

• Overview of Early Grade Reading and Math Assessments
• EGRA Coverage
• Updates to the EGRA Toolkit – Coming in February 2024
• EGRA Adherence to Eligibility Criteria for Reporting on SDG 4.1.1
• EGRA’s Potential to Expand SDG 4.1.1a Coverage
• Other USAID Assessment and Benchmarking Guidance and Work
• Recommendations and Questions
EARLY GRADE READING ASSESSMENT (EGRA) OVERVIEW

- Designed in 2008 for young learners (grades 1-4) based on the Simple View of Reading
- Refined through piloting, Toolkit updated in 2016
- Administered orally and 1:1 between a trained assessor and a child
- Adapted to and typically administered in a language that the child can use and understand
- Statistically linked to allow comparison over time
- Subtasks measure:
  - Listening comprehension and vocabulary
  - Letter naming, letter sounds,
  - Familiar words and non-words
  - Decoding
  - Decoding accuracy and fluency
  - Reading comprehension

Simple View of Reading Framework adapted by AIR, 2021
EARLY GRADE MATHEMATICS ASSESSMENT (EGMA)

OVERVIEW

- Designed for grades 1-4 in 2008
- Piloted and refined over time-Toolkit published in 2014
- Criteria for inclusion of subtasks:
  - Represent a progression of foundational skills that support proficiency
  - Research indicating predictive power—i.e., test skills related to students’ future performance
  - Common in many early grade curricula
- Administered orally and 1:1
- Adapted to local contexts and languages
- Subtasks measure:
  - Number identification
  - Number discrimination
  - Missing Number
  - Addition and Subtraction
  - Word Problems
EGRA AND EGMA ADDITIONAL QUESTIONNAIRES

• Include learner questionnaires to identify key demographic information

• Optional but common:
  • Teacher survey
  • Head teacher survey
  • School environment protocol
  • Classroom observations

• Occasionally include household surveys

• Allow countries to identify which skills learners lack, which groups are most effected, and why
EGRA DATA CAN SIGNIFICANTLY INCREASE SDG 4.1.1A COVERAGE

- 75+ countries have conducted EGRAs using more than 120 languages

- At least 23 countries have conducted nationally-representative EGRAs with Grade 2 or 3 data since 2014

- At least 3 more nationally-representative EGRAs are expected in 2024-25

- Also have 3 nationally representative EGMAs
2023/2024 EGRA CONTENT UPDATES

- New Receptive and Expressive Language Module (up to 45 items)
- Move to Decoding Accuracy at Grade 2 (about 50 words)
- Decoupling Decoding Fluency and Comprehension at Grade 3
- Allowing learners to use lookbacks for the reading comprehension questions
- Improving and Expanding Reading Comprehension (15 items)
- Ensuring EGRA Adheres to Universal Design for Assessment Practices
2023/2024 EGRA TECHNICAL UPDATES

• Clarifying the need for nationally representative data for countries that wish to report to SDG 4.1.1a
• Ensuring reporting of reliability and validity measures by domain and overall assessment
• Updating assessment confidentiality requirements
• Clarifying guidance on use of conjunctive or compensatory scoring model
• Updating benchmark-setting guidance
• Updating documentation and reporting requirements to include details on the adaptation workshop and item development/appropriateness, tool piloting, assessment psychometric properties, administration incidents, etc.
EGRA ADHERENCE TO THE ELIGIBILITY CRITERIA FOR SDG 4.1.1A REPORTING
**CRITERION 1: IS THE ASSESSMENT SUFFICIENTLY ALIGNED WITH THE MPL/GPF?**

<table>
<thead>
<tr>
<th>Level of Alignment</th>
<th>Category</th>
<th>SDG 4.1.1(a) GPF Grade 2</th>
<th>EGRA 2016-2023</th>
<th>EGRA 3.0 (2024)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Minimally Aligned</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Length</td>
<td></td>
<td>Min. score of 20 if setting &quot;meets&quot; level</td>
<td>Min. score of 45 if setting three levels</td>
<td></td>
</tr>
<tr>
<td>Domains (depth)</td>
<td>Decoding (min. 10 points)</td>
<td>Listening Comprehension (min. 5 points)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subconstructs (breadth)</td>
<td>4 of the 7 decoding and listening comprehension subconstructs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Additionally Aligned</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Length</td>
<td></td>
<td>Min. score of 20 if setting &quot;meets&quot; level</td>
<td>Min. score of 45 if setting three levels</td>
<td>Min. 60 items across listening comprehension, decoding, and reading comprehension</td>
</tr>
<tr>
<td>Domains (depth)</td>
<td>Decoding (min. 10 points)</td>
<td>Reading Comprehension (min. 5 points)</td>
<td>Decoding - 50 items Reading Comprehension - 5 items</td>
<td></td>
</tr>
<tr>
<td>Subconstructs (breadth)</td>
<td>3 of the 5 decoding and listening comprehension subconstructs</td>
<td>4 of the 5 subconstructs covered</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Strongly Aligned</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Length</td>
<td></td>
<td>Min. score of 20 if setting &quot;meets&quot; level</td>
<td>Min. score of 45 if setting three levels</td>
<td>~105 items across listening comprehension, decoding, and reading comprehension</td>
</tr>
<tr>
<td>Domains (depth)</td>
<td>Reading Comprehension (min. 10 points)</td>
<td></td>
<td>Reading Comprehension - 10-15 items</td>
<td></td>
</tr>
<tr>
<td>Subconstructs (breadth)</td>
<td>1 of the 2 reading comprehension subconstructs</td>
<td></td>
<td>At least 1 of the 2 subconstructs covered</td>
<td></td>
</tr>
</tbody>
</table>
## CRITERION 1: IS THE ASSESSMENT SUFFICIENTLY ALIGNED WITH THE MPL/GPF?

<table>
<thead>
<tr>
<th>MPL requirements</th>
<th>EGRA Toolkit Guidance</th>
</tr>
</thead>
</table>
| **01** MPL: When listening to slightly longer texts, they make simple inferences | - **Task**: Listening comprehension subtask/Receptive and expressive language module  
- **Guidance**: 2024 revision to include between 12-45 questions (5 items in the 2016 guidance) |
| **02** MPL: Accurately read aloud | - **Task**: Accuracy subtask  
- **Guidance**: 2024 revision to focus on accuracy with ~50 words (fluency was linked to comprehension in 2016 guidance) |
| **03** MPL: Retrieve explicit information | - **Task**: Reading comprehension subtask  
- **Guidance**: 2024 revision to require 10-15 questions (5 items in 2016 EGRA guidance) |
### CRITERION 2: HAVE THE ITEMS BEEN REVIEWED QUALITATIVELY AND QUANTITATIVELY?

<table>
<thead>
<tr>
<th>Criteria for Reporting on SDG 4.1.1</th>
<th>EGRA Toolkit Guidance</th>
</tr>
</thead>
</table>
| **01** Is there evidence that the items in the assessment have been reviewed qualitatively? | • Teachers and curriculum, language, and assessment experts are engaged in the tool *adaptation* process  
• Pilots conducted by teachers & trained enumerators |
| **02** Is there evidence that the items in the assessment have been reviewed quantitatively? | • Items difficulty/facility and item discrimination will be required to be reported in EGRA 3.0 Toolkit  
• Item discrimination must be greater than 0.2 |
| **03** Have items been reviewed to ensure appropriateness for relevant population subgroups? | • Assessment follows UDA principles and can be adapted for learners with disabilities  
• Adaptations consider contextual realities  
• Pilots are conducted with relevant subgroups |
| **04** Test development process is documented in detail | • EGRA guidance will provide a job aide aligned with Standards 4.7 - 4.10 to meet this criterion |
CRITERION 3: IS THE SAMPLE OF LEARNERS THAT TOOK THE ASSESSMENT REPRESENTATIVE?

<table>
<thead>
<tr>
<th>Criteria for Reporting on SDG 4.1.1</th>
<th>EGRA Toolkit Guidance</th>
</tr>
</thead>
</table>
| **01** Was the assessment administered to the whole cohort or a sample? | ● Nationally-representative sampling guidance will be provided in the job aide in the EGRA toolkit  
  ● Will include sampling when multiple assessments are used to cover different areas/languages |
| **02** Were any subgroups of the population excluded? | ● EGRA generally does not include OOS children/youth  
  ● Goal: to ensure inclusion of learners with disabilities  
  ● Guidelines around age of students? |
| **03** Is the margin of error 5% or less - 95% confidence level? | ● Sampling requirements do include this guidance in the EGRA Toolkit |
| **04** Was the MDES calculated to ensure comparison over time | ● EGRA Toolkit guidance has always required power analysis be conducted and MDES be appropriate to allow comparisons over time |
CRITERION 4: IS THERE EVIDENCE THE ASSESSMENT WAS ADMINISTERED IN A STANDARDIZED WAY?

<table>
<thead>
<tr>
<th>Criteria for Reporting on SDG 4.1.1</th>
<th>EGRA Toolkit Guidance</th>
</tr>
</thead>
</table>
| **01** Were administration guides clear on the administration process? | ● Enumerator & supervisor training and reporting requirements are standardized - EGRA Toolkit and Guidance Note for Planning/Implementing EGRA  
● IRR of .80 required; supervisor checks during admin. |
| **02** Was the assessment administered in appropriate and standardized way? | ● Each supervisor and assessor team conducts daily dual-assessor checks (i.e., score the same child) to monitor scoring discrepancies  
● Supervisors have QA checklists |
| **03** Were significant incidents recorded & relevant results excluded from the outcomes? | ● EGRA reporting guidance requires standard sections in the report methodology  
● Updates will provide specific documentation guidance around administration and incidents |
CRITERION 5: ARE THE OUTCOMES OF THE ASSESSMENT SUFFICIENTLY RELIABLE?

Is the value of coefficient alpha for the assessment greater than or equal to 0.7?

- Cronbach’s alpha is required with a value of at least 0.7, though most EGRAs have a value of 0.8
- Past EGRAs have not measured alpha by subtask due some subtasks being timed; updates will require this
CRITERION 5: ARE THE OUTCOMES OF THE ASSESSMENT SUFFICIENTLY RELIABLE?

<table>
<thead>
<tr>
<th>Country</th>
<th>Grade</th>
<th>Year</th>
<th>Language</th>
<th>Sample</th>
<th>Population</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa 3</td>
<td>2</td>
<td>2015</td>
<td>English</td>
<td>1258</td>
<td>676167</td>
<td>0.82</td>
</tr>
<tr>
<td>Africa 3</td>
<td>3</td>
<td>2015</td>
<td>English</td>
<td>1224</td>
<td>627224</td>
<td>0.84</td>
</tr>
<tr>
<td>Asia 3</td>
<td>2</td>
<td>2020</td>
<td>Local</td>
<td>2692</td>
<td>177000</td>
<td>0.95</td>
</tr>
<tr>
<td>Asia 3</td>
<td>3</td>
<td>2020</td>
<td>Local</td>
<td>2718</td>
<td>166358</td>
<td>0.95</td>
</tr>
<tr>
<td>Asia 4</td>
<td>3</td>
<td>2013</td>
<td>English</td>
<td>2463</td>
<td>2300615</td>
<td>0.90</td>
</tr>
<tr>
<td>Asia 4</td>
<td>3</td>
<td>2013</td>
<td>Local</td>
<td>2463</td>
<td>2296396</td>
<td>0.91</td>
</tr>
<tr>
<td>Africa 4</td>
<td>2</td>
<td>2016</td>
<td>Kiswahili</td>
<td>2647</td>
<td>246376</td>
<td>0.92</td>
</tr>
<tr>
<td>Africa 4</td>
<td>3</td>
<td>2016</td>
<td>Kiswahili</td>
<td>670</td>
<td>186958</td>
<td>0.92</td>
</tr>
<tr>
<td>Africa 5</td>
<td>2</td>
<td>2013-2018</td>
<td>Local</td>
<td>1701</td>
<td>25812</td>
<td>0.78</td>
</tr>
<tr>
<td>Africa 5</td>
<td>3</td>
<td>2013-2018</td>
<td>Local</td>
<td>960</td>
<td>24375</td>
<td>0.79</td>
</tr>
<tr>
<td>MENA 3</td>
<td>2</td>
<td>2018</td>
<td>Arabic</td>
<td>1465</td>
<td>48509</td>
<td>0.87</td>
</tr>
</tbody>
</table>

*RTI International, 2023
CRITERION 5: ARE THE OUTCOMES OF THE ASSESSMENT SUFFICIENTLY RELIABLE?

<table>
<thead>
<tr>
<th>Criteria for Reporting on SDG 4.1.1</th>
<th>EGRA Toolkit Guidance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>01</strong></td>
<td>Is the value of coefficient alpha for the assessment greater than or equal to 0.7?</td>
</tr>
<tr>
<td></td>
<td>- Cronbach’s alpha is required with a value of at least 0.7, though most EGRAs have a value of 0.8</td>
</tr>
<tr>
<td></td>
<td>- Past EGRAs have not measured alpha by subtask due to some subtasks being timed; required in update</td>
</tr>
<tr>
<td><strong>02</strong></td>
<td>Are QA arrangements appropriate for human-scored items with agreement &gt;80%</td>
</tr>
<tr>
<td></td>
<td>- QA described on the last slide</td>
</tr>
<tr>
<td></td>
<td>- IRR of at least 80% required in piloting</td>
</tr>
</tbody>
</table>
EGRA DATA CAN SIGNIFICANTLY INCREASE SDG 4.1.1A COVERAGE BY ADDRESSING REPORTING CRITERIA FOR EXISTING DATA AND CODIFYING CRITERIA IN UPDATED TOOLKIT
### SOME DATA MAY ALREADY MEET SDG 4.1.1A CRITERIA

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Evidence required</th>
<th>Select country 1 data (2020)</th>
<th>Select country 2 data (2021)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is the assessment sufficiently aligned with MPL/GPF?</td>
<td>MPL/GPF alignment and # items</td>
<td>Decoding (57 words), reading comprehension (5 items), and listening comprehension (5 items)</td>
<td>Decoding (60 &amp; 70 word passages), reading comprehension (5 items), and listening comprehension (5 items) in local language &amp; language of instruction (also EGMA)</td>
</tr>
<tr>
<td>2. Have the items been reviewed to determine suitability for inclusion?</td>
<td>Adaptation process and participants</td>
<td>Curriculum, language experts, statisticians, teachers, EGRA trainers conducted adaptation workshop; pilot examined item difficulty, item discrimination, and bias by language, SES and disability (item &amp; subtask)</td>
<td>Use of past adaptation, same pilot statistics conducted, and pilot and assessment examined correlations across local language &amp; language of instruction, all subtasks equated &amp; comparable to past years</td>
</tr>
<tr>
<td>3. Is the sample representative?</td>
<td>Sampling frame (2 stage)</td>
<td>All government schools in country with Grade 2 stratified by region, language; students sampled by section, classroom, and sex</td>
<td>Sampled from all public and non-state schools listed in EMIS stratified by county; students sampled by grade, section, class, sex</td>
</tr>
<tr>
<td>4. Was assessment administered in a standardized way?</td>
<td>Kappa coefficient for IRR</td>
<td>Listening: 0.83; reading: 0.80; reading comprehension: 0.83</td>
<td>Listening: 0.97; reading: 0.95; reading comprehension: 0.86 (comparable for 2nd language)</td>
</tr>
<tr>
<td>5. Are the outcomes sufficiently reliable?</td>
<td>Cronbach’s alpha</td>
<td>Listening: 0.85, reading comprehension: 0.81</td>
<td>Listening: 0.87, decoding passage: 0.9, reading comprehension: 0.91</td>
</tr>
</tbody>
</table>
USAID is making other relevant updates to its indicator reporting guidance.
USAID IS ENHANCING ACCESS TO AND USE OF ASSESSMENTS AND BENCHMARKING METHODS

- **Assessment decision tree and job aides** to support decision-makers to identify the right assessment for each country’s purpose - expected February 2024

- US Foreign Assistance “F” Indicators allow use of ANY assessments and all standardized benchmarking methods

- **EGRA specific:**
  - Expanding data summaries on the EarlyGradeReadingBarometer.org
  - Permanent repository of USAID data with a growing catalogue of EGR data sets
    - data.usaid.gov → Basic Education
    - edu-links.org/resources/find-and-access-usaid-education-data
  - Templates to standardize EGRA reporting with STATA/R code and dataset templates - expected February 2024
RECOMMENDATIONS AND OUTSTANDING QUESTIONS

Recommendations

● **Stop Rules:** Accept, at a minimum, past data (but preferably future data as well) where stop rules have been used if administering agencies are able to demonstrate that the stop rules do not significantly affect results.

● **Nationally-Representative Data:** Accept data that are nationally representative as a whole, even if that means different assessments are used for different populations (e.g., stratified by language) and linked through policy linking or another method.

● **Age of Students:** Students enrolled in the grade of instruction are included in results as long as we demonstrate their inclusion does significantly affect results.

● **Acceptance of Assessment that use Classical Test Theory:** Accept use of assessments that use CTT or IRT (reporting guidance says only IRT allowed).

● **Alignment for Individually Administered Assessments:** Criteria for alignment should be the same for individually administered assessments as what is already being accepted for group-administered assessments.
RECOMMENDATIONS AND OUTSTANDING QUESTIONS

Recommendations Continued

- PAL, USAID, and UNICEF collectively recommend constituting a Technical Advisory Group (TAG) to support EGRA, MICS-FLM, and PAL-N assessment design decisions for reporting on SDG 4.1.1a (could be a reinstatement of the 4.1.1 Review Panel)
- Agree to finalizing outstanding questions by end of January 2024 for inclusion of any required changes in forthcoming EGRA Toolkit revisions

Discussion Questions

- **Number of Additional Countries Needed:** In 2024, how many more countries need to report on SDG 4.1.1a from which regions to ensure adequate coverage to upgrade SDG 4.1.1a back to Tier 1?
- **Household versus Classroom-Based Assessments:** Will UIS accept data that are nationally representative of the population of students who are enrolled in school rather than for all children nationwide?
- **Test Confidentiality:** Will UIS accept the results from tests when used versions are released publicly and not reused to report results?
**NEXT STEPS**

- Continue examining impact of the use of stop rules
- Check domain-level alphas across
- Finalize EGRA Toolkit (with guidance from GAML, TCG, UIS)
- Continue examining past EGRA data to determine which datasets might be sufficient to include, and discuss inclusion with relevant countries
- Support UIS and other stakeholders in establishing a TAG
(1) Albertha is my name.
(2) When I grow up I want to be a lawyer.
(3) I love to be in school.
(4) I love my best Su.