Deliverable #5:

Results of Pilot Test

Including Test-Retest of the Pilot Self-Administered Early Grade Reading Assessment (SA-EGRA) and Self-Administered Early Grade Mathematics Assessment (SA-EGMA) and Concurrent Validity with Traditional EGRA/EGMA

Submitted

December 16, 2022

Submitted To

Imagine Worldwide

Attn: Dr. Karen Levesque
Head of Research
Location/information follows

1080 Edgewood Ave
Mill Valley, CA 94941

E-mail:
karen.levesque@imagineworldwide.org

RTI Administrative Point of Contact

Dr. Carmen Strigel
International Education

E-mail: cstrigel@rti.org

Submitted By

Timothy Slade, Simon King, Jennifer Ryan, Karon Harden, Leah Rosenbaum, Yasmin Sitabkhan, Peggy Dubeck, and Lachezar Hristov on behalf of

RTI International
3040 East Cornwallis Road, PO Box 12194
Research Triangle Park, NC 27709-2194 USA
www.rti.org

Imagine Worldwide: Work Order #1

Contents

Deliverable #5

Executive Summary

This report summarizes the findings of an effort to develop and validate tablet-based, self-administered assessments of English-language foundational literacy and numeracy in the early grades. RTI International developed the two assessments, known respectively as the Self-Administered Early Grade Reading Assessment (SA-EGRA) and the Self-Administered Early Grade Mathematics Assessment (SA-EGMA), with the support and at the direction of Imagine Worldwide. The assessments are deemed "self-administered," because children complete the assessments independently in response to instructions and stimuli imbedded in the tablet-based software. However, adults typically supervise the organization and conduct of the assessment as well as the collection of individual data from the tablets for analysis.

The development effort took place throughout 2022. RTI's assessment experts first conducted literature reviews to identify assessment tasks that would effectively leverage the new affordances provided by the tablet-based interface. We considered a wide range of tasks before iteratively refining the list. Once the most promising candidate tasks had been developed, software developers and web designers developed a mocked-up assessment interface consisting of several screens with multiple graphical elements. A field team in Ghana conducted user testing with 40 students across three days. The findings—grounded in both direct observation of student interactions and interview questions posed to those students—informed changes to the user interface and interaction modalities. The Ghana field team conducted a five-day field test of the revised tools, assessing more than 800 students (400 students in each of grades one and three) drawn from 20 schools in two municipalities of Greater Accra. Analysis of those results informed a further round of revisions to tasks design, instructions, and items.

The field team then used the revised SA-EGRA and SA-EGMA alongside traditional English-language EGRA and EGMA to conduct pilot testing. Intended to assess the reliability and validity of the SA-EGRA and SA-EGMA instruments, the pilot test involved two key elements. First, students were assessed with both the SA- and traditional versions of either the EGRA or the EGMA. (Half the students completed the pair of literacy assessments, while the other half completed the pair of math assessments.) This enabled an evaluation of concurrent validity, as scores on both measures for the assessed subject were available for the same student. Next, students were assessed a second time, seven days after the first assessment, using the same SA-EGRA or SA-EGMA they had previously encountered. This enabled an evaluation of test-retest reliability for the self-assessment tools. This report presents the findings of this pilot testing phase.

The findings are very encouraging. Both SA-EGRA and SA-EGMA performed well on metrics of reliability and internal validity. Furthermore, student scores on each are correlated with their scores on the traditional EGRAs and EGMAs.

While the SA-EGRA tasks differ markedly from traditional EGRA tasks, the tablet-based *Spelling* task is highly correlated ($r = 0.828$) with oral reading fluency (ORF), the typical top-line metric of interest on a traditional EGRA. We conclude that the *Spelling* task may be fruitfully used to estimate generalized (aggregate) estimates of ORF proficiency levels. It lacks the precision to estimate ORF for individual students, however, and RTI does not endorse it for that task.

SA-EGMA tasks share common items with traditional EGMA tasks, but the removal of time-based constraints due to user experience concerns eliminates the ability to assess fluency. As a result, the constructs being measured by SA-EGMA and traditional EGMA are slightly different, leading to a reduced (but still acceptable) correlation across the two assessments.

We conclude the report by proposing avenues for further research that could inform the further refinement of the administration protocols, tasks, and items of the SA-EGRA and SA-EGMA.

Self-Administered Early Grade Reading Assessment (SA-EGRA) and Early Grade Mathematics Assessment (SA-EGMA) Pilot Results

1. Introduction and Background

The purpose of this activity was to evaluate the reliability and validity of the SA-EGRA / SA-EGMA. A tool's reliability is its ability to measure the desired construct with consistency. The tool should measure consistently both across time and across items. A tool's validity is the extent to which it measures the construct of interest. For the purposes of this evaluation, the constructs we were interested in measuring were early literacy and mathematics skills.

Evaluation reports—commonly used to present the results of Early Grade Reading Assessments (EGRAs) and Early Grade Mathematics Assessments (EGMAs)—emphasize the learning outcomes of the students assessed. This report is different. It is primarily concerned with the performance of the tools themselves and whether they are fit for the purpose of evaluating student learning outcomes in foundational literacy and numeracy.

This report describes how the SA-EGRA and SA-EGMA were developed; the data obtained from the pilot tests; and the instruments' psychometric properties. The conclusion presents recommendations for the use of the SA-EGRA/SA-EGMA tools as well as avenues for further development and refinement.

Institutional Review Board (IRB) guidance was sought from RTI prior to undertaking fieldwork. The IRB determined that the study was exempt from full IRB review due to being conducted in an educational setting, involving normal educational practices, and being unlikely to adversely affect the students' opportunity to learn. IRB approval was also obtained from the Ghana government.

2. Assessment Framework

Understanding the psychometric properties of the SA-EGRA/SA-EGMA tools would enable us to make changes to their constituent tasks and items that would improve overall reliability and validity. We designed the study accordingly. The next section provides a high-level overview of the key measures of interest we sought to understand.

2.1 Reliability

To determine whether a tool can produce consistent results *over time* students need to be assessed at least twice using the same tool. This is termed a "test-retest" approach. The two assessments need to be conducted relatively close in time (e.g., a few days apart) so the results are not influenced by changes in the student's actual learning. However, they must not happen too closely in time (e.g., the same day) lest the student recall their responses to the first assessment and rely upon that familiarity with the items during the second assessment. We retested students one week following the first assessment.

There are two statistical tests used for test-retest reliability. The first of these is a simple correlation using Pearson's r (the Pearson product-moment correlation), a measure of the

generalized linear association between two sets of data. [1] The second approach is a Bland-Altman[2] analysis which assesses the level of agreement between pairs of repeated measures across the spectrum of the student ability levels. For Bland-Altman analyses, a simple heuristic is to verify that <5% of observations fall outside the bounds of ±2 standard deviations (SDs).

Reliability can also be measured across tasks or items (rather than over time). If some items within a task appear to be assessing different constructs, removing or replacing them may yield a task that more consistently (reliably) assesses a single construct. Factor analysis measures how closely items within a task (or tasks within an assessment) are related to each other. Generally, the factor loadings are considered acceptable at a level of 0.3 or higher.[3] We apply factor analysis both within tasks (at the item level) and across tasks (to assess reliability of the overall tools).[4]

The items within each task can also be assessed using Item Response Theory (IRT). In classical test theory, the students being assessed are the unit of analysis. In IRT, the test itself is the unit of analysis. If the items within a task measure the same construct, they can be compared in terms of their difficulty, their ability to discriminate between students of different abilities, and their bi-serial correlation (the correlation between students' scores on the item and total scores on the task).

2.2 Validity

The validity of a tool is the degree to which it *actually* measures what it has been designed to measure. The SA-EGRA/SA-EGMA are designed to measure early grade literacy and numeracy skills respectively; the scores students achieve on each should therefore be associated with the scores they achieve on the traditional assessor-administered EGRA/EGMA.

Relative to the traditional EGRA/EGMA, the SA-EGRA/SA-EGMA introduce differences in the medium of assessment (tablet-based stimuli vs. paper stimuli) and the protocol (self-administered vs. assessor-administered) while presenting some limitations in task design (e.g., the absence of fluency measures). By adopting an approach called *concurrent validity*—having the same student complete both the traditional and SA- versions of the assessments—we are able to explore the extent to which the SA-EGRA/SA-EGMA are able to measure our constructs of interest despite these differences.

3. Tool Development Process

3.1. App Development

For the sake of brevity, this report will not revisit in full the item specifications detailed in the

[1] When applied to continuous measures, a Pearson's r >0.7 is preferred. There is no definitive rule for an acceptable threshold of correlation when using Pearson's r in the context of discrete measures with limited items, which is the situation for nearly all of the EGRA/EGMA and SA-EGRA/SA-EGMA items.

[2] Bland, Martin J., and Altman, Douglas G., 1986. "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement." The Lancet 327 (8476): 307–10. https://doi.org/10.1016/S0140-6736(86)90837-8

[3] Costello, Anna B, and Jason W Osborne. 2005. "Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis." Exploratory Factor Analysis 10 (7): 9.

[4] Because Pearson's correlation and factor analysis are both correlation-based, they have limitations when applied to discrete and binary data rather than continuous measures. For more please see Kolenikov, Stanislav, and Gustavo Angeles. 2009. "Socioeconomic status measurement with discrete proxy variables: is principal component analysis a reliable answer?" Review of Income and Wealth 55 (1): 128–65. https://doi.org/10.1111/j.1475-4991.2008.00309.x

earlier document titled *Draft Assessment Specifications and Prototype (APK) Ready for Field Testing: Summary of the Task and Item Development for the Pilot Self-Administered Early Grade Reading Assessment (SA-EGRA) and Self-Administered Early Grade Mathematics Assessment (SA-EGMA)* (RTI International, 2022).[5] The sections below provide a high-level description of the tasks and skills assessed by the SA-EGRA and SA-EGMA as well as the overall process by which the tools were refined.

a. SA-EGRA tasks and literacy skills assessed

Letter Sound Recognition

The *Letter Sound Recognition* task assesses the student's knowledge of letter sound-symbol correspondence at the phoneme level. Unlike on the traditional EGRA, in which the student reads letters sounds aloud from a grid of 100 for up to one minute, on the SA-EGRA this task contains 10 items in a receptive, multiple-choice format. For each item the student is presented with 5 written letters on the screen and hears the sound associated with one of them. The student then taps the letter that corresponds to the sound that s/he hears. The student may tap a button to hear the sound again as many times as s/he wants before answering. Distractor letters are chosen based on either their phonological or visual similarity to the target letter. Between the target letters and the distractors, the task incorporates nearly the full range of English phonemes. On the SA-EGRA this task is untimed and therefore measures only accuracy of recognition, not fluency.

Spelling

The *Spelling* task further assesses the student's ability to apply their knowledge of letter-sound correspondences and common spelling patterns to encode words. The task format is a dictation in which the student is asked to transcribe 10 words given orally. For the initial prompt, the student hears the word in isolation, hears it used in the context of a sentence, then hears it in isolation twice more. S/he then spells (types) out the word using a virtual alphabet strip. The student can tap a button to rehear the word as many times as desired. The student is given partial credit for partially correct answers.

Silent Reading Comprehension

The *Silent Reading Comprehension* task assesses the student's ability to understand a text (i.e., written words presented in the meaningful context of a story). In the SA-EGRA, the student reads a slightly longer (190-word) passage to him/herself and then answers 10 multiple-choice questions about it one at a time. In order to mitigate the limitations of short-term memory, the student is allowed to refer back to the passage at will while answering the questions, though s/he cannot go back to previously answered questions. The questions and answer options (though not the reading passage itself) are also presented both in written and oral form.

Language Proficiency

Finally, the *Vocabulary* and *Syntax* tasks assess these two important aspects of the student's proficiency in English. Assessing language proficiency alongside reading is one way to tease out whether any reading comprehension difficulties are due to low language proficiency, a distinction that is especially important in contexts where students are learning to read in languages that they

[5] Interested readers may request access to this report.

don't speak at home.

The *Vocabulary* task captures data about the breadth of the student's receptive knowledge of English words. The task presents the student with 20 multiple choice items, each with one real word and three pseudowords. The student is prompted to "Tap the word that you know the best." The real word is the only option that the student could actually know. If the student does not know the target word, s/he will resort to guessing. All of the pseudowords conform to English phonology and orthography and therefore look like possible words. All of the words are presented in both oral and written form, and the student can tap a button to hear the word again as many times as desired. The oral stimulus is to mitigate the possible effects of low reading ability; the written stimulus helps the student hold all four options better in their short-term memory while deciding among them. Since English is presumably a second language for this population, the target real words were selected from among the 1000 most frequent words in English.

The *Syntax* task assesses the student's knowledge of how words are put together to make meaning in English. The task presents the student with 10 sentence pairs that use the same words but differ only in word order; one makes sense and the other does not, and the student is instructed to tap the sentence that makes sense. The sentences are presented both orally and in written form for the same reasons as mentioned above, and the student can tap a button to hear them again as many times as desired. The sentences contain only very common words to minimize the potential of a student's limited vocabulary being a constraint. The items focus on a variety of common English sentence structures that second language learners would be expected to know as 3rd grade students having received three years of oral English language instruction, ranging from very simple to slightly more complex.

b. SA-EGMA tasks and mathematical skills assessed

As with the SA-EGRA, the SA-EGMA does not currently assess fluency. However, the SA-EGMA assesses the same mathematical skills as the traditional EGMA through the following tasks: number identification, number discrimination, missing number identification, addition, subtraction, and word problems.

Number Identification

The *Number Identification* (aka "number ID") task evaluates students' ability to connect the spoken name for a number (e.g., "three") with its symbolic representation (e.g. "3"). Unlike the traditional EGMA, the SA-EGMA speaks aloud the name of a number (e.g., "three") and asks students to enter the corresponding symbolic representation (e.g., "3"). To account for this lengthier interaction, the number of items was reduced from 20 to 12 during development and the time restriction (necessary to assess fluency) was removed.

Number Discrimination

The *Number Discrimination* (aka *number comparison*) task evaluates students' ability to discern between quantities, represented symbolically as numbers, and to identify the largest number (e.g., that 58 is larger than 49 and 32). This subtest also assesses students' place-value skills by presenting pairs in which the larger number has a smaller ones or tens digit than the smaller number (e.g., 534 vs 287 vs 199). Based on the results of the field test, the SA-EGMA used at the pilot stage presents students with triples of numbers instead of pairs (e.g. 8 vs 12 vs 6 instead of 8 vs 12) to reduce the chance of students selecting the correct response purely by chance.

Missing Number

The *Missing Number* task evaluates students' ability to identify a missing element in a sequence of numbers (e.g., 14, 15, __, 17). The 10 test items are identical to the traditional EGMA. Instead of asking students to say the name of the missing number aloud, as in the traditional EGMA, the SA-EGMA instead asks students to enter the missing number using an array of digits. The format enables visual verification not afforded by the traditional format. That is, students can see the complete pattern in the self-assessment version whereas they only say the number aloud in the traditional version.

Addition and Subtraction

The *Addition* and *Subtraction* tasks assess students' addition and subtraction skills using single-digit (Level 1) and two-digit (Level 2) numbers. *Addition Level 2* and *Subtraction Level 2* are only administered to students who give at least one correct response to the corresponding Level 1 assessment. Due to the increased interaction time, the number of items in Level 1 tasks was reduced from 20 to 13 during development. The number of items in *Addition Level 2* and *Subtraction Level 2* remains the same at five items each. All *Addition* and *Subtraction* tests are untimed and thus do not assess fluency.

Word Problem

The *Word Problem* task evaluates students' ability to understand an arithmetic problem described in narrative form (e.g., "Three students are on a bus. Two more students join them. How many students are on the bus now?") in a way that allows them to operate on and solve the problem (e.g., 3 + 2 = ___, counting on from 3, etc.). Students may use pencil and paper as they work on each problem, though only the final answer as entered into the tablet is recorded. The six word problems are identical to those on the traditional EGMA. While the tablet cannot use verbal and non-verbal cues to check students' understanding as administrators of the traditional EGMA are trained to do, students may repeat the problem narrative as many times as they wish.

c. Stages of Assessment

Development of data collection tools ideally proceeds through three main stages.

1. **User testing:** this stage occurs during initial and subsequent renderings of a software product. The product is tested iteratively with small user groups or individuals, with feedback collected both via direct observation and explicit questions.
2. **Field testing:** this stage assesses the initial performance of a tool, with an objective to refine the tool and address issues prior to piloting.
3. **Pilot testing:** in this stage, a tool is fully assessed for its psychometric properties, with a particular focus on reliability and validity.

The SA-EGRA/SA-EGMA development proceeded accordingly. The field test included analysis of multiple-choice items, the assessment duration, protocols, and overall construct validity. (The analysis of field test data is summarized in the document *Preliminary findings: Pilot*.[6]) It is the analysis of the data collected during the pilot testing phase that forms the basis of the findings detailed in this evaluation report.

[6] Interested readers may request access to the summary deck.

d. User testing

User testing included both ad-hoc and structured components. For the ad-hoc components, RTI team members who had children of early-elementary school ages observed as their children interacted with the app and completed the assigned tasks (or attempted to). These experiences provided early insight into task duration concerns, as well as the layout and scale of the on-screen interactive elements. For the structured component, a field team in Ghana tested the SA-EGRA and SA-EGMA user interfaces with 40 students from both urban and rural schools over a three-day span (June 14-16, 2022).

The team made minor changes to the user interface (its look, feel, and interactive elements) in response to the user testing feedback. For instance, an additional interactive element was included in order to separate "help" functions (such as stating and reiterating task instructions) from "content" functions (such as stating or reiterating an item or prompt). Also, task instructions referred to one of the visual elements—an animated character resembling a Tangerine who provided additional guidance when touched—as "the tangerine". In response to feedback that students in the target area were unfamiliar with tangerines, the team updated the audio script to replace references to "the Tangerine" with references to "the fruit".

e. Field testing

The initial field test was conducted from August 1, 2022 through August 5, 2022 with 441 grade one and 429 grade three students at 20 schools in the Adenta and Weija-Gbawe Municipalities in Greater Accra, Ghana. During the field test, students were assessed using either the SA-EGRA or the SA-EGMA instrument. Data was collected at a single timepoint, and the findings were used to refine the administration protocol, tasks, and items. The main focus of the findings was to:

- Update the app's rendering to address issues either observed during administration or evident from data analysis.
- Observe how the students interacted with the app to identify any required changes to the assessment protocol, whether in the classroom or on the app.
- Assess tasks and task items for internal consistency. Adapt, change or remove tasks or items that do not perform to expectations.
- Measure the average duration for each assessment and determine whether the assessments should be shortened to mitigate issues such as fatigue that could threaten the tools' validity.

The summary of findings and action items from the field test are summarized in ***Annex A***.

f. Pilot testing

The pilot test was conducted from September 28, 2022 through October 11, 2022. It included both a concurrent-validity component (with the same student completing both a traditional EGRA or EGMA and its self-administered counterpart) and a test-retest component (with each student completing the SA-EGRA or SA-EGMA a second time 7 days after they were initially assessed).

- To assess the tools' performance across all ability ranges, we sought an even spread (i.e., a uniform distribution) of student abilities. The concurrent validity approach allowed us to closely monitor oral reading fluency scores on the traditional EGRA and make daily adjustments to the student selection procedures to ensure we obtained the desired range of student abilities.

2. Pilot test findings

a. Student literacy outcomes

Exhibit 1: SA-EGRA Pilot Average Learning Outcomes, by task

SA-EGRA Task	Average Percent Score
Syntax[7]	77.5%
Letter Sounds	72.6%
Vocabulary	54.7%
Spelling	44.0%
Silent Reading Comprehension	32.5%

The student literacy average percent scores should be viewed in context of the sampling approach used to select students. The grade 3 students were selected purposefully each day in order to develop an approximately uniform distribution of student performance along an oral reading fluency scale. Consequently, the average percent scores from the pilot (***Exhibit 1***) do not represent the average population performance, because the sample was purposefully drawn to represent a range of abilities and was not a representative sample.

The pattern of average percent scores from this study can be usefully compared to a traditional EGRA. As with a traditional EGRA, students score most proficiently in Syntax (roughly akin to an EGRA's Listening Comprehension task; 77.5%) and Letter Sounds (72.6%). The most challenging tasks were Spelling (the only productive / non-multiple choice task; 44.0%) and the higher-order literacy skill of comprehension (32.5%).

b. Student mathematics outcomes

While students who were assessed in literacy were explicitly sampled to ensure a roughly uniform spread across literacy levels, students who were assessed in mathematics were randomly sampled from the available Grade 3 students in each of the schools. In general, the outcomes of the SA-EGMA tasks—presented in ***Exhibit 2***—follow similar patterns to traditional EGMA tasks.

Exhibit 2: SA-EGMA Pilot Average Learning Outcomes, by task

SA-EGMA Task	Average Percent Score
Number Identification	65.3%
Number Discrimination	83.6%
Missing Number	59.7%
Addition Level 1	87.5%
Addition Level 2	64.3%
Subtraction Level 1	70.9%
Subtraction Level 2	53.7%
Word Problems	38.3%

The only outlier from this pattern is the first task, *Number Identification*. As the SA-EGMA tasks are always presented to students in the same order, this is the first task students encountered in the self-administered format. This pattern was first encountered in the field test, where student responses were indicative of potential confounding between learning the input features of the tablet—including by entering long sequences of random multiple-digit numbers—and the actual assessment task. This behavior was seen across all student ability levels, with some students scoring highly on higher order tasks, but very poorly on *Number Identification*. Steps were taken between the field test and the pilot to account for these issues by introducing a short game prior to this first task, instructing students to play with the data input system for a short time. While this reduced the under-performance issue somewhat, it still has not wholly eliminated it; these behaviors will need to be considered during future deployments. *Number Identification* is a fundamental mathematical skill; we strongly recommend exploring modifications to item construction or administration protocol to mitigate these issues.

[7] The tasks grouped under the *Syntax* rubric were binary-choice items; the other forced-choice tasks included four or five answer options. The baseline rate of correctness due to lucky guesses is thus slightly different.

c. Time taken to complete assessment

Exhibit 4: Mean Assessment Durations for SA-EGRA and SA-EGMA, by time point

Assessment	Mean (mins.)	Standard Error
SA-EGRA (t_1)	41.5	±0.95
SA-EGRA (t_2)	37.8	±3.08
SA-EGMA (t_1)	49.7	±2.36
SA-EGMA (t_2)	41.7	±2.86

The time required for a student to complete an assessment is an important consideration when considering its use. While the assessment duration has logistical implications, student fatigue can also affect performance. ***Exhibit 3*** at right summarizes the mean duration of students' assessments at both time points.

The mean duration of students' SA-EGRA assessments did not reduce significantly from t_1 to t_2. The case of SA-EGMA was different, however; students completed the assessment in substantially fewer minutes the second time they encountered it.[8] We recommend analysis of task- and item-level durations to understand the impact on overall assessment time of factors like (a) the time students spent playing with the tablet during the *Number Identification* task, or (b) the relative length (or susceptibility to skipping) of item-level audio prompts.

a. Internal Consistency

SA-EGRA Tasks

Exhibit 3: Factor Analysis Loadings for SA-EGRA task scores

SA-EGRA Task Percent Score	Factor 1 Loadings
Syntax	0.619
Letter Sounds	0.550
Vocabulary	0.731
Spelling	0.874
Silent Reading Comprehension	0.622

Here we present a summary of the tools' internal consistency; the detailed analyses are provided in ***Annex B***. We assessed the overall internal consistency for the SA-EGRA using factor analysis of the task percent scores. This provides an opportunity to assess if the tasks were measuring the same latent construct. Our analysis, summarized in ***Exhibit 4*** at right, found strong factor loadings into a single construct ranging from 0.550 (*Letter Sounds*) to 0.874 (*Spelling*). Given that acceptable factor scores should be 0.3 or more, the internal consistency at the summary level for the SA-EGRA is excellent.

The SA-EGRA tasks were also assessed for internal consistency at the item level for each task using both factor analysis and Item Response Theory (IRT). We present the analysis for the *Syntax* task in ***Exhibit 5*** below and discuss it as an example. For the sake of brevity, the comparable tables for the other tasks are included in ***Annex B***. The discussion here will be limited to key points in summary form.

[8] A deeper analysis of item- and task-level variance is beyond the scope of this report but may provide insights that could inform revisions to protocol or item selection.

Exhibit 5: Item Factor Analysis and IRT for the Syntax task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	*Difficulty*	*Bi-serial Correlation*
1	0.207	0.60	0.72	0.48
2	0.324	0.63	0.73	0.57
3	0.196	0.72	0.60	0.52
4	0.139	0.42	0.81	0.43
5	0.109	0.72	0.58	0.43
6	0.221	0.52	0.76	0.49
7	0.325	0.51	0.81	0.54
8	0.486	0.40	0.89	0.60
9	0.372	0.40	0.87	0.53
10	0.311	0.36	0.88	0.52

For IRT to work well the items being analyzed should explain the same construct. In this case, items 1, 3, 4, 5 and 6 have factor loadings less than 0.3, suggesting that these items either explain a different construct or may need adjustment. Such inconsistencies can arise, for example, when even high ability students are not able to correctly answer the items. For example, a deeper exploration (omitted here for space, but available upon demand in the study's analytical tables) showed that students who correctly answered item #8 were only able to answer item #5 (which had a low factor loading) 61% of the time, whereas the same students who answered item #8 correctly were able to correctly answer item #9 (which had a higher factor loading) over 92% of the time. The conclusion is that an item might need adjustment if students are more likely to guess or misunderstand the item.

The IRT item-level analysis for *Syntax* starts with *Discrimination*. This reports the difference between the proportions of high and low scorers answering an item correctly. A discrimination score of over 0.2 helps an item to contribute towards the measurement of variable student ability. The discrimination scores for the *Syntax* items are acceptable, ranging from 0.36 (Item #10) to 0.72 (Item #5). Item *Difficulty* is a simple calculation of the proportion of students who correctly answered that item. It ranges from 0 (no student answered correctly) to 1 (all students answered correctly). For an assessment designed specifically to measure the variability of student skills, *Difficulty* scores should ideally range from 0.2 to 0.8. Half of the *Syntax* items fall within this range, with the remainder of the items exhibiting *Difficulty* scores of over 0.8. *Syntax* was the highest-scoring task in general; these *Difficulty* results should be interpreted in context of the literacy skills being assessed and the difficulty of the other tasks. The bi-serial correlation is the Pearson correlation between responses to a particular item and scores on the overall task. It ranges from -1 to 1, and strong positive correlations are desirable. The bi-serial correlations for the *Syntax* items are acceptable, ranging from 0.48 to 0.6. In sum, while exploration of whether the easier items (*Difficulty* score > 0.8) should be updated could be undertaken as part of iterative improvement, the *Syntax* task performed well and can be deployed with confidence as currently designed.

Internal consistency for *Letter Sounds* was acceptable overall, although item #7 demonstrated a concerningly low factor loading (0.0018) (***Exhibit B-3*** *of* ***Annex B***). The *Discrimination* for item #7 was 0.16 and its *Difficulty* was 0.17. Upon inspection, the correct response was the letter sound

/n/, while 77% of the students actually responded */m/*.[9] Should opportunities for further analysis arise, a further review of this item and the other items with factor loadings less than 0.3 (#s 4, 5, 8, and 9) would be valuable to understand why these items do not identify as well with the common latent trait as do the other *Letter Sound* items. Potential explanations for these results given the multiple-choice format may include poorly performing distractors. However, it is also possible that an item is well-constructed and this performance is instead highlighting an issue with the instruction and mastery of that particular item.

The internal consistency for the *Vocabulary* task is shown in ***Exhibit B-4*** of ***Annex B***. The *Discrimination* and *Difficulty* scores are acceptable, ranging from 0.4 to 0.78 and 0.29 to 0.81 respectively. We recommend further analysis to understand the low factor loadings for a few items (#s 12, 16, 20), as this might inform item-level substitution ahead of a subsequent deployment.

For *Silent Reading Comprehension*, while a few items (#s 1, 7, and 8) display low factor loading and merit further review, *Discrimination* and *Difficulty* are within acceptable ranges for all items.

The *Spelling* task has different item characteristics than the other SA-EGRA tasks. Other tasks present the student with a binary or multiple choice and score it as correct or incorrect. The *Spelling* task requires students to actively produce text and the scoring mechanism awards partial credit for incorrect responses. IRT analysis—which primarily deals with binary items that are fully correct or fully incorrect—is therefore not suitable for this task. The item factor analysis (***Exhibit B-6*** *of* ***Annex B***) is excellent, with factor loads being between 0.725 and 0.853, easily surpassing the 0.3 threshold. Many different factors contribute to the strength of this task. From a psychometric perspective, this task (unlike the others) provides minimal opportunity for correct guessing; is able to discriminate between different skill levels by awarding partial credit; and can yield an aggregate score between 0 and 63 (a wide continuous measure). As a consequence, the *Spelling* task is excellent for capturing the variability of students' skill levels.

At the task level, factor 1 loadings exceeded the heuristic threshold of 0.3 by a substantial margin. Overall, our analysis indicates that the SA-EGRA tasks discussed here were internally consistent. We look forward to further scrutinizing the performance of specific individual items as additional data becomes available from other deployments of the tool.

[9] Following similarly low performance during the field test the team scrutinized the audio prompt and concluded it was potentially ambiguous. A higher-quality recording of the audio prompt substituted at the pilot stage in order to reduce the ambiguity. The continued high degree of error warrants scrutiny, although it may simply highlight that teaching and mastery of the difference between */n/* and */m/* may bear additional emphasis.

SA-EGMA Tasks

As with the SA-EGRA, we assessed the overall internal consistency of the SA-EGMA using factor analysis of the task percent scores. Our analysis, summarized in ***Exhibit 6***, found moderately strong factor loadings onto a single construct ranging from 0.356 (*Number Identification*) to 0.652 (*Missing Number*). Given that acceptable factor scores should be 0.3 or more, the internal consistency of the tasks for the SA-EGMA is good.

Exhibit 6: Factor Analysis Loadings for SA-EGMA task scores

Task Percent Score	Factor 1 Loadings
Number Identification	0.355
Number Discrimination	0.438
Missing Number	0.652
Addition	0.544
Addition Level 2	0.533
Subtraction	0.462
Subtraction Level 2	0.557
Word Problems	0.534

IRT techniques are a poor fit for free-response item types. Nearly all SA-EGMA tasks with the exception of *Number Discrimination* were presented in a free-response format; as a result, we did not perform IRT analyses of the SA-EGMA. Item-level internal consistency analyses for the SA-EGMA were limited to factor analysis.

Exhibit 7: Item Factor Analysis for the *Missing Number* task

Item	Factor Analysis
1	0.330
2	0.276
3	0.375
4	0.004
5	0.531
6	0.139
7	0.488
8	0.414
9	0.554
10	0.179

Exhibit 7 at left presents the item-level factor analysis for the *Missing Number* task. While it was the task with the highest internal consistency at the task level, its item-level characteristics are similar to those of the other tasks.

Six of the ten items surpass the target threshold (a loading of 0.3 on the first factor). Several items, however, have concerningly low loadings, including items #4, 6, and 10. Items 4 and 6 are the only two which involve three-digit numbers. Item 10 requires repeated addition of 5, but unlike item 8 (which does so as well), the target value is not a multiple of 5. (e.g., `1, [ ], 11, 16` rather than `5, 10, [ ], 20`.)

Item-level analysis of the other SA-EGMA tasks reveals similar patterns. Roughly half of the items in each task fall below the threshold of 0.3 for first-factor loading. Some of those items have borderline-acceptable loadings in the range of 0.25-0.29; however, most tend to be substantially lower (ranging from 0.20 to as low as 0.004).

As with SA-EGRA items that have low factor loadings, further analysis and investigation is warranted before deciding that items with low factor loadings should be revised; other issues may be at play. For instance, *Number Identification*—the SA-EGMA task with the lowest factor loadings—was also the first task the students faced. Within *Number Identification*, the first item had the lowest loading. (Indeed, it was among the items on the entire SA-EGMA [across all tasks] with the lowest first-factor loading.) The item required students to enter the numeral "2" after hearing it read aloud; that it appears to load onto a different factor than the other items may speak more to the students' familiarity with the assessment modality than with the item proper. (Indeed, the first three items were three of the four lowest-loading items in the task; all of them then clustered on factor 2 with loads of 0.49 or above.)

At the task level, factor 1 loadings substantially exceeding 0.3 indicate that the SA-EGMA tasks are internally consistent. Item-level analysis suggests that certain items may warrant further scrutiny to determine whether substitution is appropriate.

b. Test-Retest Reliability

SA-EGRA Tasks

We assessed test-retest reliability using Pearson's correlation to report the generalized relationship between student scores assessed at the two timepoints. Pearson's correlation is generally used when reporting linear associations of continuous variables. When applied to variables with discrete outcomes and a relatively limited number of items, as is the case with the SA-EGRA tasks, lower Pearson's correlations are to be expected.[10] ***Exhibit 8*** at right reports the correlations for student scores on the same task at the two timepoints; the graphs depicting the individual students' scores are presented in ***Annex C, Exhibit C-2***.

Exhibit 8: Pearson's Correlation for the SA-EGRA Test-Retest

SA-EGRA Task Percent Score	Correlation
Syntax	0.569
Letter Sounds	0.598
Vocabulary	0.721
Spelling	0.904
Silent Reading Comprehension	0.660

The *Spelling* task demonstrates the strongest positive correlation at the two time points. While the task's continuous scoring and wide range counteract some of the limitations that Pearson's correlation encounters with the other tasks, and may contribute to the very strong correlation, as discussed previously *Spelling* also demonstrated excellent international consistency, which may contribute to this result. (See ***Exhibits 4*** and **B-6**). The other tasks demonstrate correlations in the range of approximately 0.5 to 0.7.

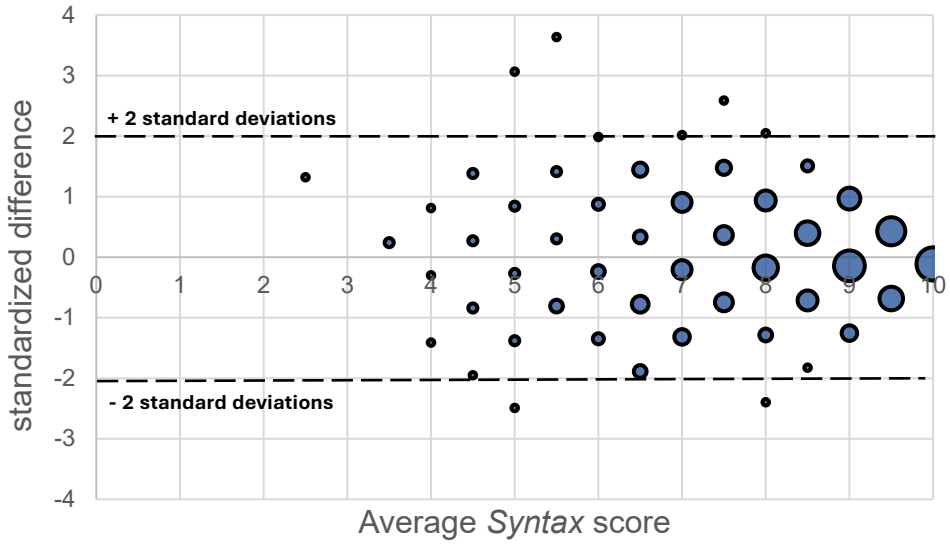
As mentioned, the limited number of items and discrete scoring of the non-*Spelling* tasks poses a challenge for interpretation of Pearson's correlation. We therefore additionally include Bland-Altman analyses. Bland-Altman analyses present the level of agreement between pairs of repeated measures across the spectrum of the student ability levels, rather than the relationship of two sets of results as with Pearson's correlation.

To illustrate the principle, ***Exhibit 9*** below presents the Bland-Altman graph for students' scores on the *Syntax* task. The student's average score across the two timepoints is plotted along the horizontal axis. The standardized difference between the student's scores at the two timepoints (score at t_2 – score at t_1) is plotted along the vertical axis.[11] For normally distributed data, we would expect 95% of standardized differences to be within ±2 standard deviations (represented on the graph by the two dotted lines). This is indeed the case for all of the SA-EGRA tasks, including *Syntax*. The infrequency with which students scored outside the ±2 SD band is an indication that the tasks exhibit strong test/retest reliability. The Bland-Altman graphs for the other tasks are included in ***Annex C***.

[10] As noted earlier, there is no definitive threshold above which Pearson's *r* as applied to a limited number of discrete items would be preferred.

[11] If a student scores higher at t_1, they will be represented by a point below the horizontal zero axis. The score difference is then standardized (i.e., converted to standard deviations).

Exhibit 9: Bland-Altman Graph for SA-EGRA Test-Retest, *Syntax* task



The Bland-Altman plots and Pearson correlation plots indicate that the SA-EGRA tasks demonstrate very good test-retest reliability (***Annex C***). Notably, a small percentage of students scored very well at one timepoint and very poorly at the other. (The points representing these students tend to cluster near the center of the horizontal axis and beyond the ±2 bounds. They are present both above and below the 0 line, indicating that t_2 was not uniformly the higher-scoring timepoint for these students.) We suspect this phenomenon is likely driven less by the tool itself and more by how the student approached the assessment at one of the timepoints; regardless, it warrants further investigation. While it may not be possible to fully eliminate the underlying cause, a modification to the administration protocol or instructions may mitigate it.

SA-EGMA Tasks

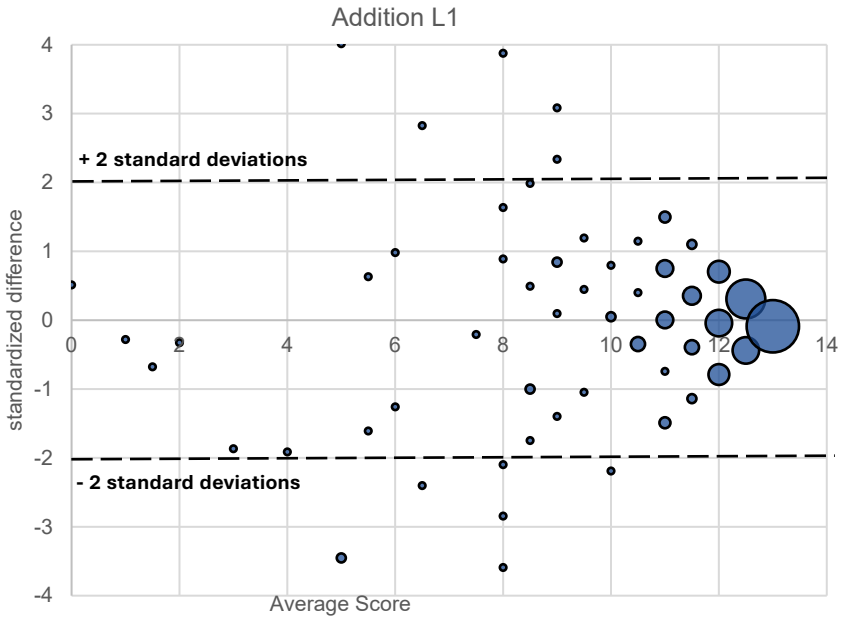
We conducted the same test-retest reliability analyses for SA-EGMA as for SA-EGRA. **Exhibit 10** at right reports the correlations for student scores on the same task at the two timepoints; the graphs depicting the individual students' scores are presented in **Annex C**, **Exhibit C-5**.

Exhibit 10: Pearson's Correlation for the SA-EGMA Test-Retest

SA-EGRA Task Percent Score	Correlation
Number Identification	0.587
Number Discrimination	0.650
Missing Number	0.811
Addition Level 1	0.499
Addition Level 2	0.424
Subtraction Level 1	0.572
Subtraction Level 2	0.400
Word Problems	0.753

None of the SA-EGMA tasks involved either continuous measures or very large numbers of items. As with the SA-EGRA, caution should be taken in over-interpreting the Pearson's correlations; we again complement the analysis with Bland-Altman plots. The plot for *Addition Level 1*, presented in ***Exhibit 11*** below, displays similar patterns as observed with the SA-EGRA tasks. It is also generally illustrative of the other SA-EGMA tasks.

Exhibit 11: Bland-Altman Graph for SA-EGMA Test-Retest, *Addition Level 1* task



The points representing the substantial majority of the students fall within the expected range of ±2 standard deviations. The fanning effect along the vertical center line, indicating students who scored very well at one time point and very poorly at the other, is very pronounced. However, the number of students who scored very similarly—and very well—at both timepoints, as seen from the rightward-pointing arrowhead pattern centered along the horizontal axis, is substantial as well. As with SA-EGRA, it was very rare for students to have widely divergent scores at t_1 and t_2. Among the SA-EGMA tasks, *Addition Level 1* has the greatest proportion of such outliers falling beyond the expected range, at 5.4%; for each of the other tasks fewer than 4.3% of students fell outside that range.

We conclude from our Bland-Altman analyses that both the SA-EGRA and SA-EGMA tasks exhibit a high degree of reliability across repeated administrations. This lends confidence that each tool indeed provides a stable measure of an underlying attribute. The next section discusses the evidence supporting whether the underlying attributes in question are early literacy and mathematics skills as measured by the traditional EGRA and EGMA.

c. Construct Validity

SA-EGRA Tasks

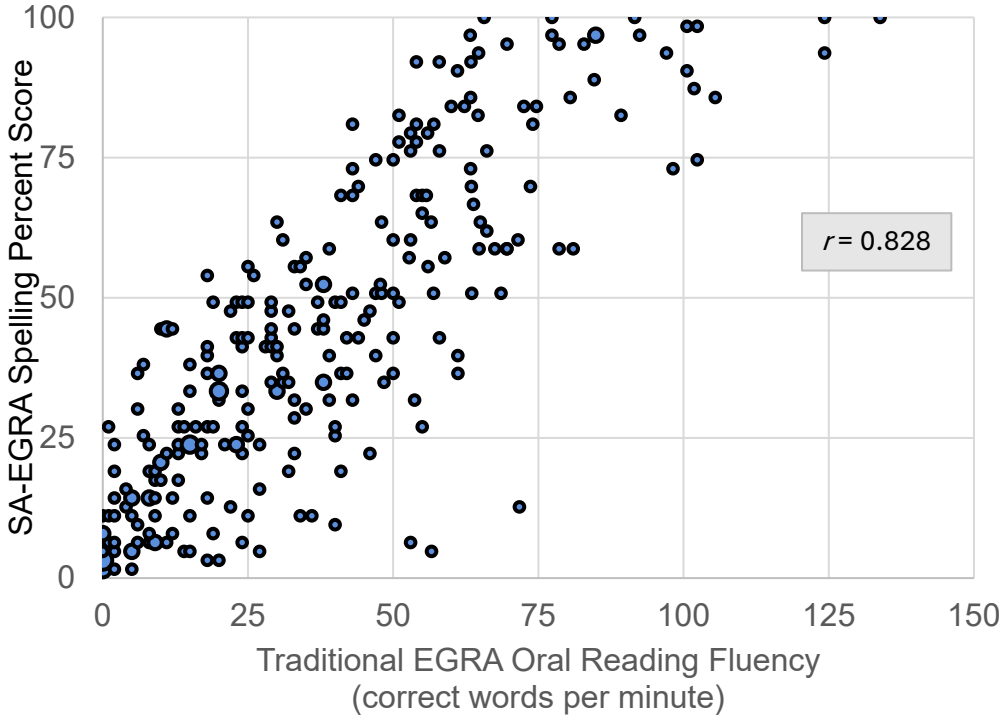
Construct validity is typically used to compare a new tool to an existing tool that has been shown to measure the same construct. A high level of association between the new and existing tools suggests that they indeed measure the same construct. However, the administration modality of the SA-EGRA differs substantially from that of the traditional EGRA; comparing mostly multiple-choice tasks (on the SA-EGRA) with oral responses to grids of items (as with many traditional EGRA tasks) is not comparing like-for-like. Additionally, there will always be interest to know if the SA-EGRA is comparable to the traditional EGRA's oral reading fluency (ORF) task due to the latter's

extensive use as measure of program impact and reporting against the United Nations' SDG 4.1.1a.[12]

We therefore sought a method of exploring SA-EGRA's concurrent validity with the traditional EGRA that would be likely to meet the needs of the SA-EGRA's target user base. We decided to explore whether the SA-EGRA can be used as a proxy measure for ORF. One option would be to create a composite SA-EGRA score combining the task percent scores weighted according to relative importance derived from expert judgment. However, given the excellent performance of the *Spelling* task, we also elected to assess how well it is associated with ORF on the traditional EGRA. This process is called statistical linking.

A student who spelled every item correctly would have earned 63 points on the *Spelling* task. (Each item was worth $n+2$ points, where n was the number of letters in the word.) The percent score was simply the student's score divided by 63. The correlation scatterplot comparing the scores on the SA-EGRA *Spelling* task and the traditional EGRA's ORF task is shown in ***Exhibit 12***, below.

Exhibit 12 SA-EGRA Spelling Percent Score vs. ORF on the Traditional EGRA



The correlation of $r = 0.828$ indicates a strong positive linear association between the two tasks. While this association is strong, it is also important to explore the predictive ability of the model. For example, students who scored 60% on *Spelling* recorded ORF scores of between 31 and 71 correct words per minute. This suggests that while this statistical linking of SA-EGRA *Spelling* results with traditional EGRA ORF scores could be used for generalized equivalent findings (such as population estimates), it lacks the precision to provide a 1:1 mapping between SA-EGRA and

[12] United Nations. 2019. "SDG Indicators." SDG Indicators. Retrieved December 2022. https://unstats.un.org/sdgs/metadata/?Text=&Goal=4&Target=4.1.

traditional EGRA for individual students. Should the opportunity for further analysis arise, we recommend exploring whether techniques such as structural equation modeling (SEM) may enable the development of a composite score that could be more highly predictive of either population-level or individual students' ORF scores.

Exhibit 13: Pearson's Correlation for Generalized Concurrent Validity of the SA-EGMA and Traditional EGMA

SA-EGMA Task Percent Score	Correlation
Number Identification	0.188
Number Discrimination	0.388
Missing Number	0.633
Addition Level 1	0.456
Addition Level 2	0.436
Subtraction Level 1	0.571
Subtraction Level 2	0.383
Word Problems	0.654

SA-EGMA Tasks

As discussed previously, there is a very tight coupling at both the task and item levels between the SA-EGMA and the traditional EGMA. As a result, using Pearson's correlation for generalized assessment of concurrent validity is a more appropriate method than was the case for the SA-EGRA and traditional EGRA. ***Exhibit 13*** below presents the Pearson's correlation for each task.[13]

The most notable feature of the SA-EGMA's task-level correlations with the traditional EGMA is how widely they range. The *Number Identification* is very weakly correlated ($r = 0.188$), while *Missing Number* ($r = 0.633$) and *Word Problems* ($r = 0.654$) are much more strongly so. Given how little the items themselves differed across assessments, this suggests that differences in administration modality may have a substantial influence over students' math scores. Pending further exploration of this phenomenon and potential revisions to either items or protocol, we do not recommend attempting to use SA-EGMA scores to predict traditional EGMA scores.[14]

3. Conclusions

The primary aim of the effort which culminated in this report was to develop *tablet-based tools* that students could *self-administer* and that would *effectively and reliably* assess their *foundational literacy and numeracy skills*. We strove to develop tools that would have high concurrent validity with the well-known and widely used "traditional" EGRA and EGMA. We believe the effort has been successful.

Literacy:

- SA-EGRA tasks have acceptable internal consistency. Some individual items with low loadings on factor 1 may merit further examination for future iterations of the SA-EGRA.
- SA-EGRA tasks have good test-retest reliability. The Pearson correlation for students' test scores across two timepoints was acceptable. Bland-Altman plots revealed both the tasks' strong performance for students across the full range of abilities and the infrequency of a given student recording highly divergent scores.
- The SA-EGRA *Spelling* task has good construct validity with respect to the traditional

[13] As noted earlier, there is no definitive threshold above which Pearson's r as applied to a limited number of discrete items would be preferred.

[14] While *Missing Number* may have been used in various contexts as a country-level benchmarking task, we would recommend against this application. While the *Missing Number* task relates to a core mathematical skill (pattern identification), it is only one way of measuring that. *Number Discrimination* and *Addition* are tasks that represent fundamental math concepts (*magnitude* and *operations*, respectively) while sidestepping some of the issues with the *Missing Number* task (such as students' potential unfamiliarity with the task, sensitivity to formatting/presentation of the items, etc.).

EGRA's oral reading fluency task. It could be used to produce generalized (population) estimates of ORF proficiency levels. However, we do not recommend it be used to estimate ORF of individual students because it lacks precision.

Mathematics:

- SA-EGMA tasks have acceptable internal consistency. As with the SA-EGRA, some individual items have low loadings on factor 1. Further examination of those items may suggest potential revisions to the items or the protocol, at the risk of diverging more from the traditional EGMA item bank.
- SA-EGMA tasks have acceptable test-retest reliability. The Pearson correlation for students' test scores across two timepoints was acceptable. The patterns of the Bland-Altman plots suggest a more pronounced clustering of high- and low-ability students' scores than in the SA-EGRA, but instances of a given student recording highly divergent scores across the timepoints remained acceptably rare.
- The design of the SA-EGMA tasks drove changes to the underlying constructs compared to the traditional EGMA. For instance, as the SA-EGMA tasks are no longer subject to a time constraint, they do not provide a measure of fluency. Similarly, the post-field test update of the *Number Discrimination* task from a 2-way comparison to a 3-way comparison changes the underlying construct.
- Generalized concurrent validity between the SA-EGMA and the traditional EGMA is acceptable. While Pearson correlations at the task level range widely, this is to be expected in light of the changes to the constructs being assessed. Further investigation of the underlying factors could improve confidence in the findings and may suggest ways SA-EGMA results could also be used to estimate traditional EGMA results.

4. Recommendations

Considerations for Deployment of SA-EGRA and SA-EGMA

The RTI team has reasonable confidence about deploying the current versions of the SA-EGRA and SA-EGMA in English-speaking contexts. We believe the current research demonstrates that assessments conducted with the SA-EGRA and SA-EGMA are valid, reliable, and effectively target foundational literacy and numeracy skills respectively. As with all assessment tools, iterative improvement is possible. As evidence about the tools' performance in varying contexts accumulates, minor changes to specific items and protocols may be appropriate and commendable.

Where logistical considerations and/or funding permit, there are a few steps organizations could take that might yield improved results. For instance, some of the scores on the SA-EGMA's early tasks are suggestive of students exploring interactions with the tablets, possibly in a playful manner. Where an organization has an existing relationship to a school in which these tools will be deployed, it may be that giving students opportunities to play with the tablets—whether via the "brain break" bubble-popping game added to the SA-EGMA, or a "sandbox" task where students can freely enter text of their choosing—could mitigate some of the degradation in data quality observed in those early tasks.

Furthermore, while the pilot testing stage of this study included a test-retest component, we do not know whether repeated use of the same items will result in so-called "learning effects" which undermine the utility of the assessments. When using traditional EGRAs, generating new test

forms through randomization of grid tasks is a simple and effective way to combat such learning effects. However, given the SA-EGRA's substantially lower total item counts, a simple re-randomization of items may not provide the same benefit. (The SA-EGMA, like the traditional EGMA, is designed so the difficulty of successive items increases gradually within a task.) This could be empirically evaluated: a subsequent study could assess both the effect on student performance of repeated exposure to re-randomized forms of the assessment and the comparability of alternate test forms with different items.

Avenues for Further Research

The field testing and piloting of the SA-EGRA and SA-EGMA generated a rich set of data. The analyses presented in this report, while sufficient for high-level decisions about the suitability of the tools for further deployment and/or adaptation into additional languages, are not exhaustive. Should resources be available, secondary analyses could be pursued that may prove fruitful in guiding refinement of the SA-EGRA and SA-EGMA tools. Below we present some suitable directions of future analysis and research.

SA-EGRA:

1. Further analysis of SA-EGRA data using structural equation modeling (SEM) could support the development of a composite score for the SA-EGRA on a more robust basis than *a priori* expert judgment. SEM could reveal whether dropping certain tasks or items could improve overall construct validity.
2. Some revisions were made to SA-EGRA task formats following the field test, such as attempting to alleviate so-called "yes-bias" in the *Syntax* task by replacing a paired-scoring approach to dyads of items with simply presenting the sentences side by side and asking the student to choose among them. While the current analysis demonstrated that the *Syntax* task performed well, a deeper dive into the data could reveal the benefits and drawbacks of the two approaches.

SA-EGMA:

As discussed, the overall performance of the SA-EGMA was strong. Nonetheless, some findings warrant further exploration.

1. A deeper dive into the factor analysis may help understand the low factor loadings of many items. This may also suggest further modifications to scoring or administration protocols. For instance, the *Number Identification* displayed both the lowest concurrent validity with the traditional EGMA and the lowest task-level factor loadings. How might those results change if the first three items—the first three non-practice items faced by the student being assessed, each of which is loading heavily onto factor 2—were excluded from scoring? Alternatively, might it be preferable to add (and then potentially not score) three additional "new" items to serve as a warm-up, so the *Number Identification* items to be scored continue to match the traditional EGMA items?
2. We suspect the comparatively low Pearson's correlation between the SA-EGMA and traditional EGMA—despite sharing common items—is rooted in the SA-EGMA assessing slightly different constructs than the traditional EGMA. Deeper analyses of the factor loadings may serve to lend credence to or refute this hypothesis.

Both:

1. An analysis of task- and item-level durations could be informative for revising administration protocols or identifying items whose performance (e.g., as discriminators) is poorly aligned with the time students spend on them.

These SA-EGRA and SA-EGMA tools are but the first iteration of a form of assessment that is novel in many low- and middle-income country contexts. The tools' overall performance was encouraging, and we believe the evidence supports a conclusion that it is appropriate to deploy them tools in their current form. That said, improvements are always possible. Minor changes to administration protocol, task instructions, or item selection may yield improvements in validity and reliability. We recommend remaining alert to opportunities to iteratively revise these tools as evidence accumulates about their performance in various contexts.

Annexes

Annex A. Tabular summary of field test data suggesting instrument modifications

Exhibit A-1: Field Test Recommendations for modifications to the SA-EGRA

Task	Recommendations
Letter Sounds	Refine poor-performing distractors on 4 items. Review audio issue on one item.
Oral Reading Rate (Moving Windows Format)	Needs more development to be a consistent measure of oral reading fluency. Drop for pilot. Remove "moving windows" and convert to a static, one-page, full-view format. Reduce length of passage.
Oral Reading Rate Filter (Comprehension)	Reduce number of items.
Silent Reading Comprehension	Refine distractors in two items, remove two items.
Vocabulary	Evidence of affirmation bias. Change format to multiple choice.
Spelling	Performed well. Reduce items from 16 to 10.
Syntax	Evidence of affirmation bias. Refine task such that correct and incorrect sentences are presented in pairs.

Exhibit A-2: Field Test Recommendations for modifications to the SA-EGMA

Task	Recommendations
Number Identification	Introduce a short game at the beginning of the assessment to allow students to get used to typing input, hopefully to reduce issues in input for this first task. End task after 4 incorrect items in a row.
Number Discrimination	Change from two response options to three. End task after 4 incorrect items in a row.
Missing Number	End task after 4 incorrect items in a row.
Addition	End task after 4 incorrect items in a row.
Addition Level 2	End task after 4 incorrect items in a row. Skip task if student scores zero on Level 1.
Subtraction	Introduce Bubble Pop Game as a break for students before this task, to keep frustration low. End task after 4 incorrect items in a row.
Subtraction Level 2	End task after 4 incorrect items in a row. Skip task if student scores zero on Level 1.
Word Problems	End task after 4 incorrect items in a row.

Annex B. SA-EGRA and SA-EGMA Internal Consistency Analysis

SA-EGRA

The first factor loadings for SA-EGRA are displayed below. As discussed earlier, a factor loading of 0.3 or higher is desirable. The factor loadings range from 0.62 (*Syntax* task percent score) to 0.87 (*Spelling* task percent score). The task-level internal consistency of the SA-EGRA is excellent.

Exhibit B-1: Factor Analysis Loadings for SA-EGRA task scores

SA-EGRA Task Percent Score	Factor 1 Loadings
Syntax	0.6191
Letter Sounds	0.5498
Vocabulary	0.7305
Silent Reading Comprehension	0.6224
Spelling	0.8740

Exhibit B-2: Item Factor Analysis and IRT for the *Syntax* task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
1	0.207	0.6	0.72	0.48
2	0.324	0.63	0.73	0.57
3	0.196	0.72	0.6	0.52
4	0.139	0.42	0.81	0.43
5	0.109	0.72	0.58	0.43
6	0.221	0.52	0.76	0.49
7	0.325	0.51	0.81	0.54
8	0.486	0.4	0.89	0.6
9	0.372	0.4	0.87	0.53
10	0.311	0.36	0.88	0.52

Exhibit B-3: Item Factor Analysis and IRT for the *Letter Sounds* task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
1	0.3574	0.52	0.80	0.57
2	0.3562	0.39	0.87	0.53
3	0.3424	0.47	0.84	0.57
4	0.2065	0.62	0.66	0.49
5	0.2327	0.37	0.85	0.46
6	0.4159	0.45	0.86	0.59
7	0.0018	0.16	0.17	0.17
8	0.1375	0.54	0.67	0.46
9	0.1334	0.58	0.66	0.46
10	0.3616	0.32	0.88	0.52

Exhibit B-4: Item Factor Analysis and IRT for the *Vocabulary* task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
item 1	0.305	0.68	0.52	0.54
item 2	0.226	0.57	0.63	0.48
item 3	0.231	0.55	0.52	0.49
item 4	0.266	0.54	0.74	0.5
item 5	0.198	0.61	0.51	0.46
item 6	0.278	0.59	0.73	0.52
item 7	0.335	0.69	0.45	0.57
item 8	0.370	0.75	0.66	0.59
item 9	0.394	0.78	0.48	0.61
item 10	0.291	0.63	0.56	0.55
item 11	0.331	0.69	0.58	0.56
item 12	0.104	0.41	0.29	0.35
item 13	0.212	0.58	0.43	0.48
item 14	0.227	0.4	0.81	0.47
item 15	0.331	0.62	0.68	0.56
item 16	0.160	0.42	0.72	0.42
item 17	0.205	0.63	0.43	0.47
item 18	0.297	0.67	0.58	0.54
item 19	0.235	0.65	0.54	0.5
item 20	0.156	0.43	0.77	0.4

Exhibit B-5: Item Factor Analysis and IRT for the *Silent Reading Comprehension* task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
1	0.013	0.21	0.21	0.2
2	0.337	0.68	0.43	0.58
3	0.264	0.59	0.38	0.52
4	0.367	0.69	0.35	0.61
5	0.543	0.72	0.32	0.64
6	0.529	0.75	0.39	0.6
7	0.123	0.39	0.24	0.41
8	0.041	0.14	0.26	0.17

Exhibit B-6: Item Factor Analysis for the *Spelling* task

Item	Factor Analysis
1	0.768
2	0.735
3	0.796
4	0.853
5	0.852
6	0.815
7	0.801
8	0.855
9	0.811
10	0.725

SA-EGMA

The first factor loadings for SA-EGMA are displayed below. As discussed earlier, a factor loading of 0.3 or higher is desirable. The factor loadings range from 0.355 (*Number Identification* task percent score) to 0.652 (*Missing Number* task percent score). The task-level internal consistency of the SA-EGMA is good, albeit less strong than the SA-EGRA.

Exhibit B-7: Factor Analysis Loadings for SA-EGMA task scores

Task Percent Score	Factor 1 Loadings
Number Identification	0.355
Number Discrimination	0.438
Missing Number	0.652
Addition	0.544
Addition Level 2	0.533
Subtraction	0.462
Subtraction Level 2	0.557
Word Problems	0.534

Exhibit B-8: Item Factor Analysis for the *Number Identification* task

Item	Factor Analysis
1	0.052
2	0.104
3	0.157
4	0.252
5	0.109
6	0.216
7	0.403
8	0.215
9	0.118
10	0.234

Exhibit B-9: Item Factor Analysis for the *Number Discrimination* task

Item	Factor Analysis
1	0.156
2	0.294
3	0.146
4	0.304
5	0.274
6	0.091
7	0.463
8	0.428
9	0.373
10	0.357

Exhibit B-10: Item Factor Analysis for the *Missing Number* task

Item	Factor Analysis
1	0.330
2	0.276
3	0.375
4	0.004
5	0.531
6	0.139
7	0.488
8	0.414
9	0.554
10	0.179

Exhibit B-11: Item Factor Analysis for the *Addition Level 1* task

Item	Factor Analysis
1	0.329
2	0.380
3	0.373
4	0.132
5	0.009
6	0.475
7	0.187
8	0.466
9	0.252
10	0.266

Exhibit B-12: Item Factor Analysis for the *Subtraction Level 1* task

Item	Factor Analysis
1	0.599
2	0.604
3	0.497
4	0.240
5	0.267
6	0.337
7	0.195
8	0.420
9	0.044
10	0.261

11	0.205	11	0.354
12	0.247	12	0.066
13	0.227	13	0.260
14	0.233	14	0.227
15	0.133	15	0.301
16	0.199	16	0.100
17	0.386	17	0.246
18	0.408	18	0.235

Exhibit B-13: Item Factor Analysis for the *Word Problems* task

Item	Factor Analysis
1	0.548
2	0.083
3	0.318
4	0.250
5	0.561
6	0.488

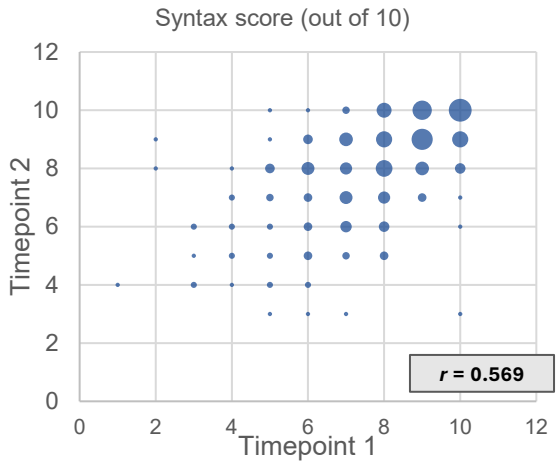
Annex C. Test-Retest Reliability: Pearson's Correlation and Bland-Altman Plots

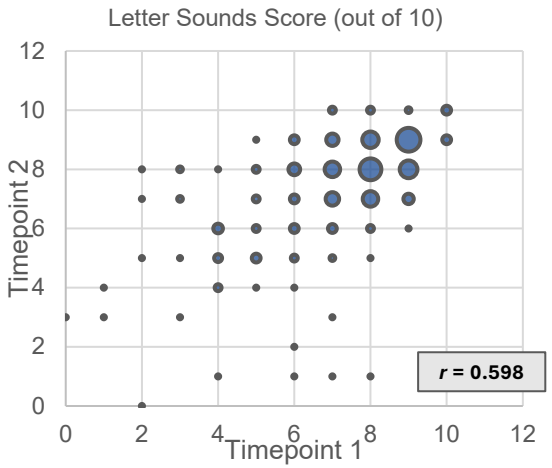
SA-EGRA

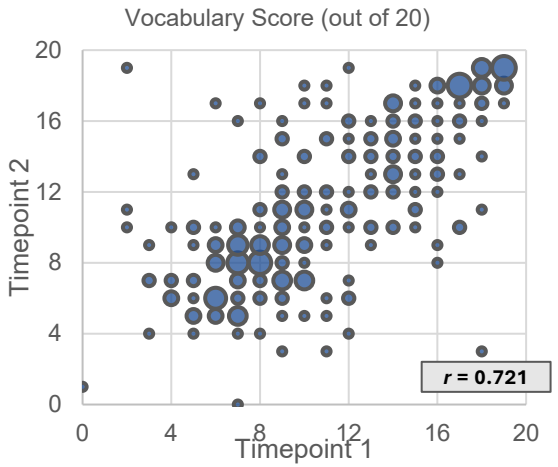
Exhibit C-1: Pearson's Correlation for the SA-EGRA Test-Retest

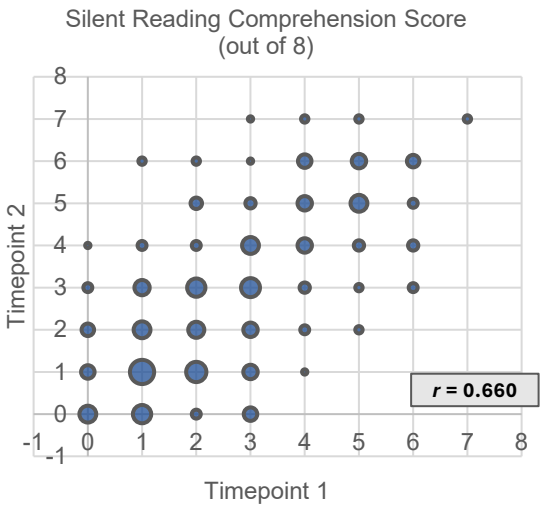
SA-EGRA Task Percent Score	Correlation
Syntax	0.569
Letter Sounds	0.598
Vocabulary	0.721
Spelling	0.904
Silent Reading Comprehension	0.660

Exhibit C-2: SA-EGRA Pearson's Correlation Generalized Test-Retest Reliability









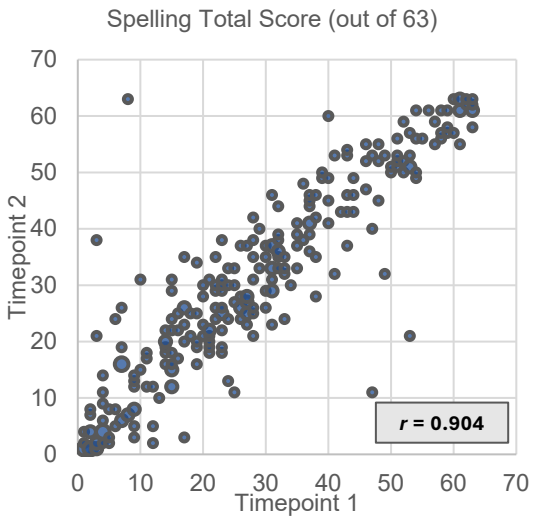
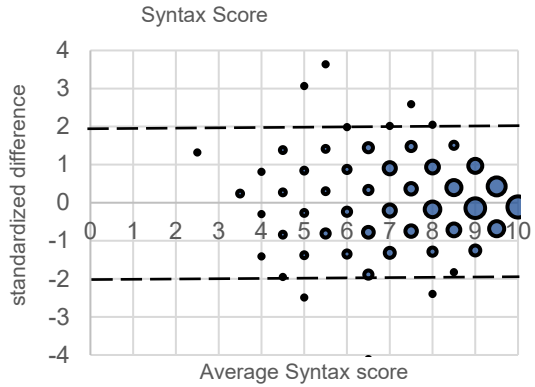
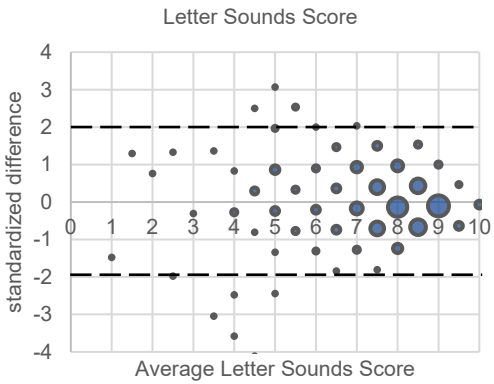
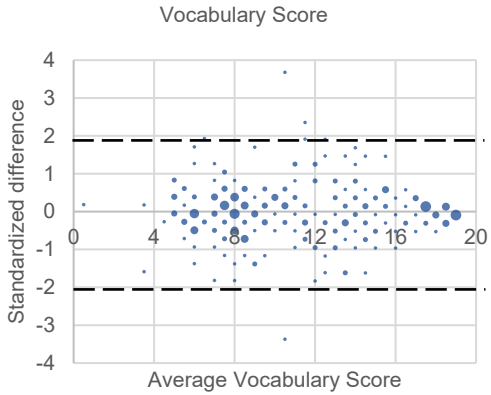


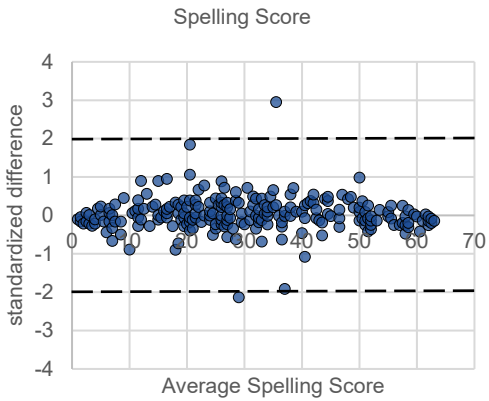
Exhibit C-3: SA-EGRA Bland-Altman Plots of Test-Retest Reliability, by task








Silent Reading Comprehension score
Standardized difference
4
3
2
1
0
-1
-2
-3
-4
0
1
2
3
4
5
6
7
8
Average Silent Reading Comprehension score

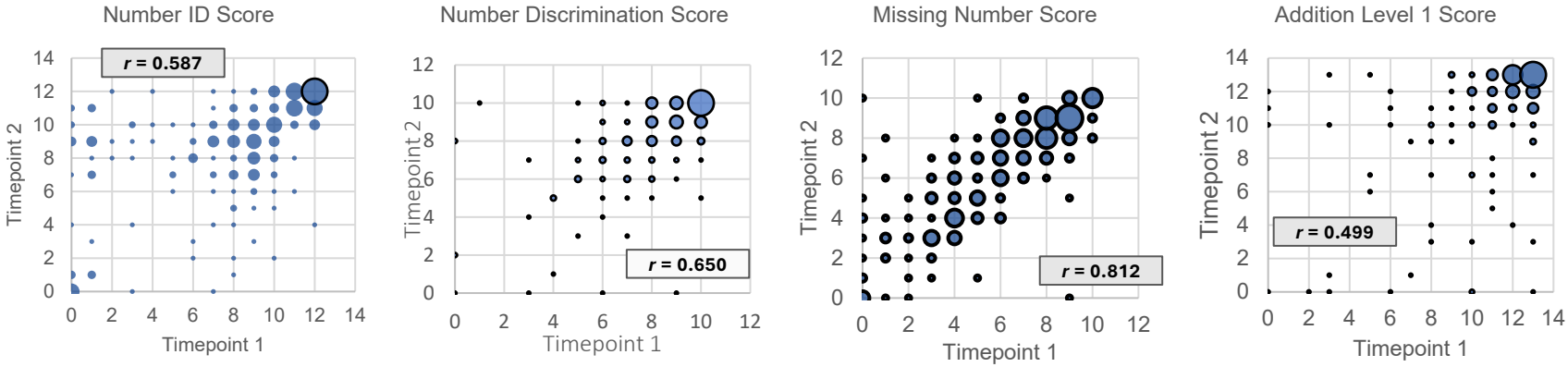

Spelling Score
standardized difference
4
3
2
1
0
-1
-2
-3
-4
0
10
20
30
40
50
60
70
Average Spelling Score

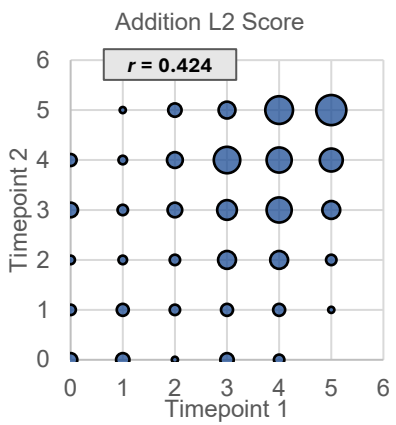
SA-EGMA

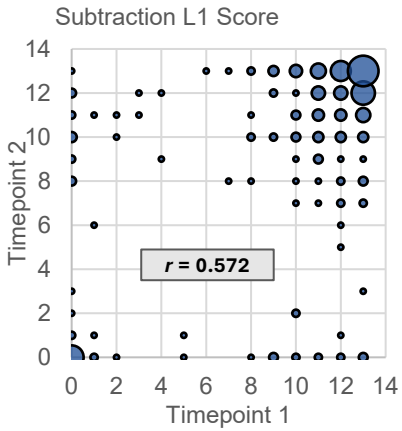
Exhibit C-4: Pearson's Correlation for the SA-EGMA Test-Retest

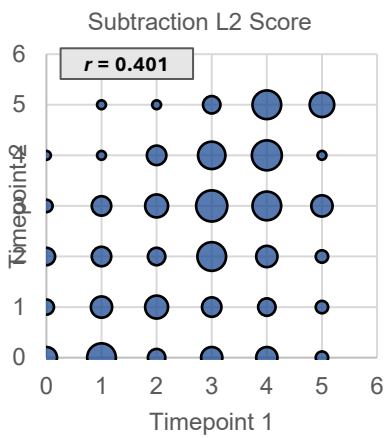
SA-EGRA Task Percent Score	Correlation
Number Identification	0.587
Number Discrimination	0.650
Missing Number	0.811
Addition Level 1	0.499
Addition Level 2	0.424
Subtraction Level 1	0.572
Subtraction Level 2	0.400
Word Problems	0.753

Exhibit C-5: SA-EGMA Pearson's Correlation Generalized Test-Retest Reliability




Addition L2 Score
r = 0.424
Timepoint 2
Timepoint 1
0 1 2 3 4 5 6


Subtraction L1 Score
r = 0.572
Timepoint 2
Timepoint 1
0 2 4 6 8 10 12 14


Subtraction L2 Score
r = 0.401
Timepoint 2
Timepoint 1
0 1 2 3 4 5 6

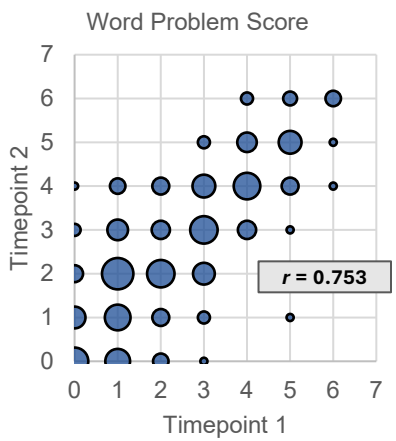
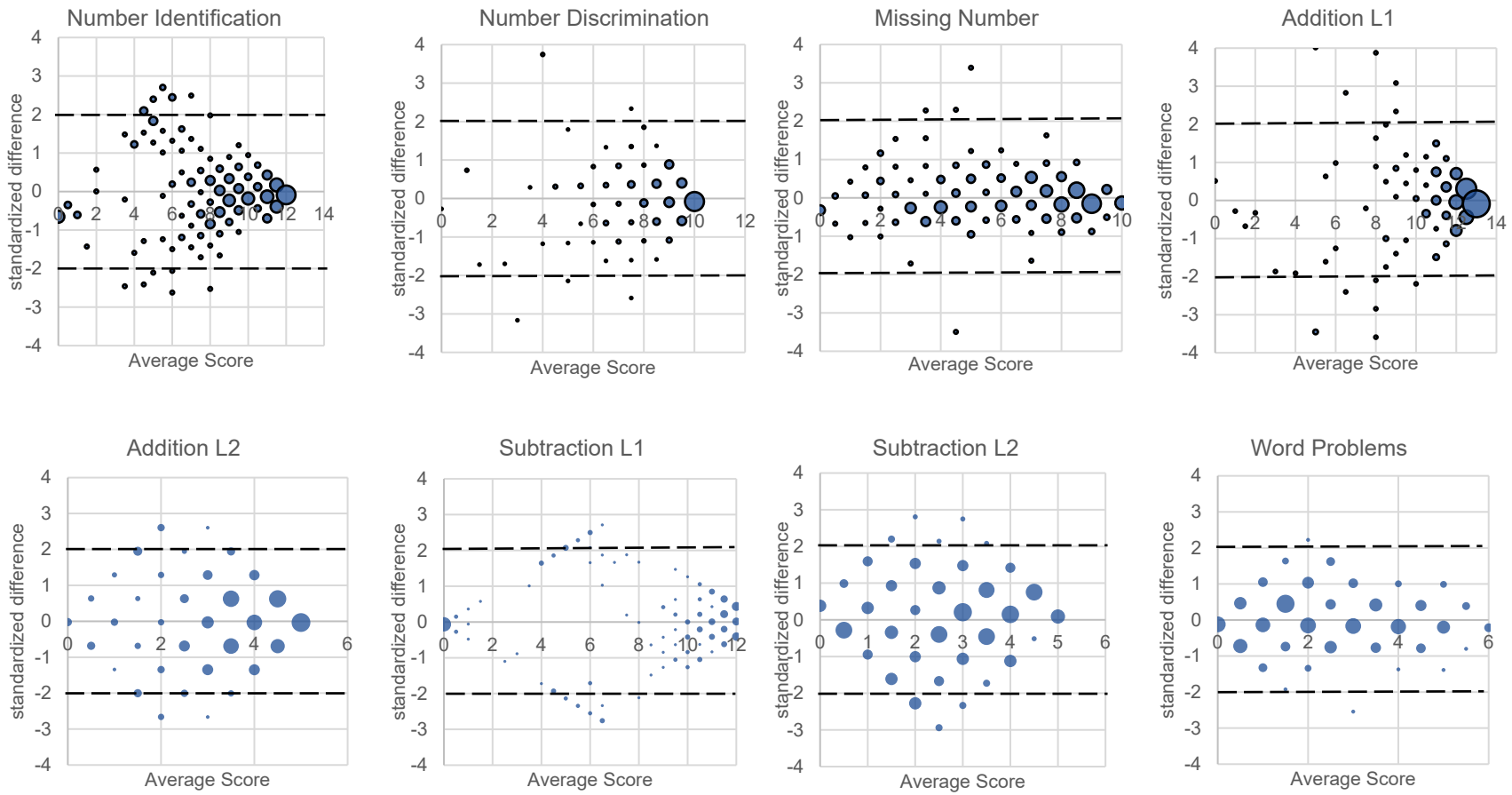

Word Problem Score
r = 0.753
Timepoint 2
Timepoint 1
0 1 2 3 4 5 6 7

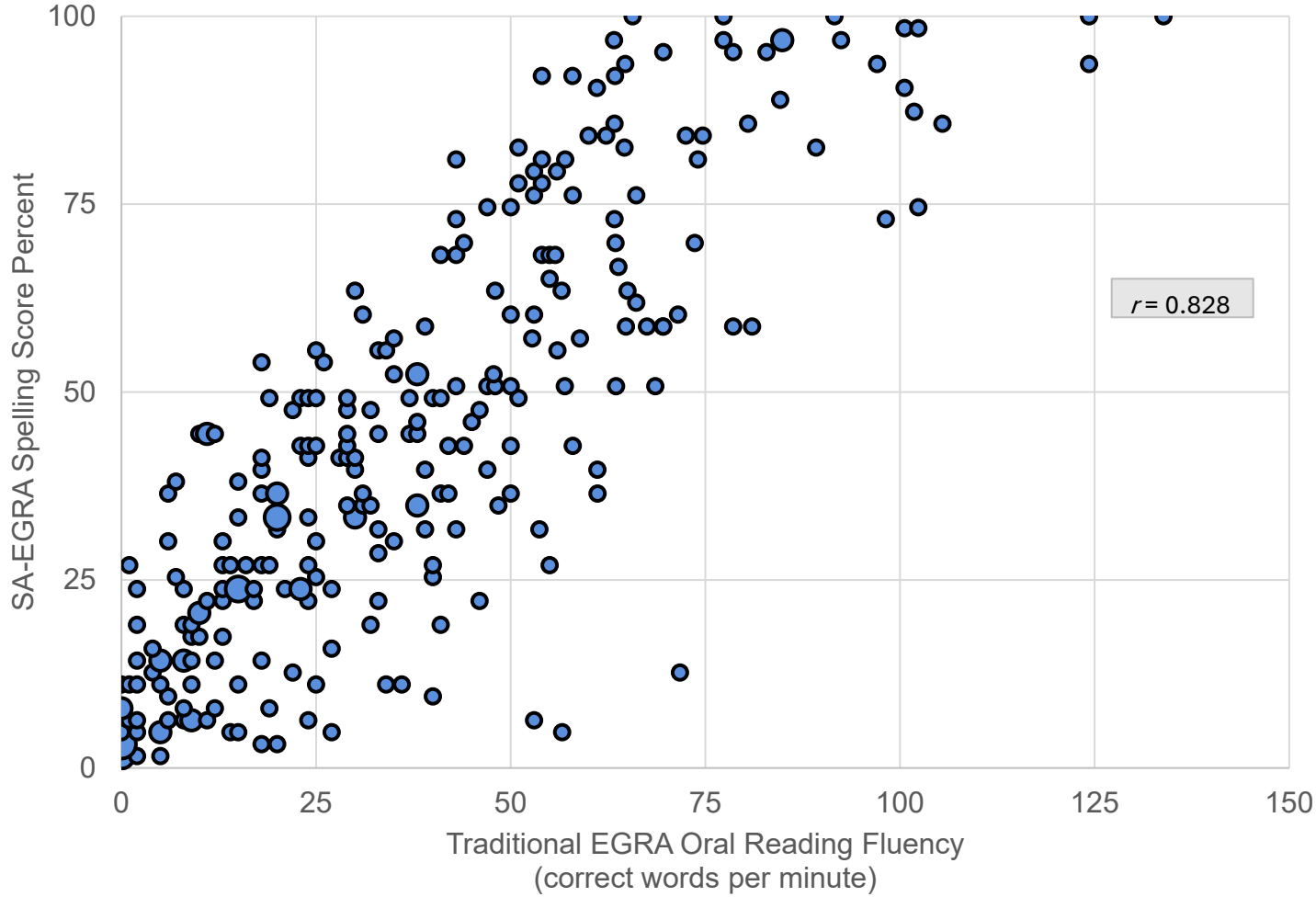
Exhibit C-6: SA-EGMA Bland-Altman Plots of Test-Retest Reliability, by task



Annex D. SA-EGRA Construct Validity and SA-EGMA Concurrent Validity

SA-EGRA

Exhibit D-1 SA-EGRA Spelling Score Percent score vs. Traditional EGRA Oral Reading Fluency

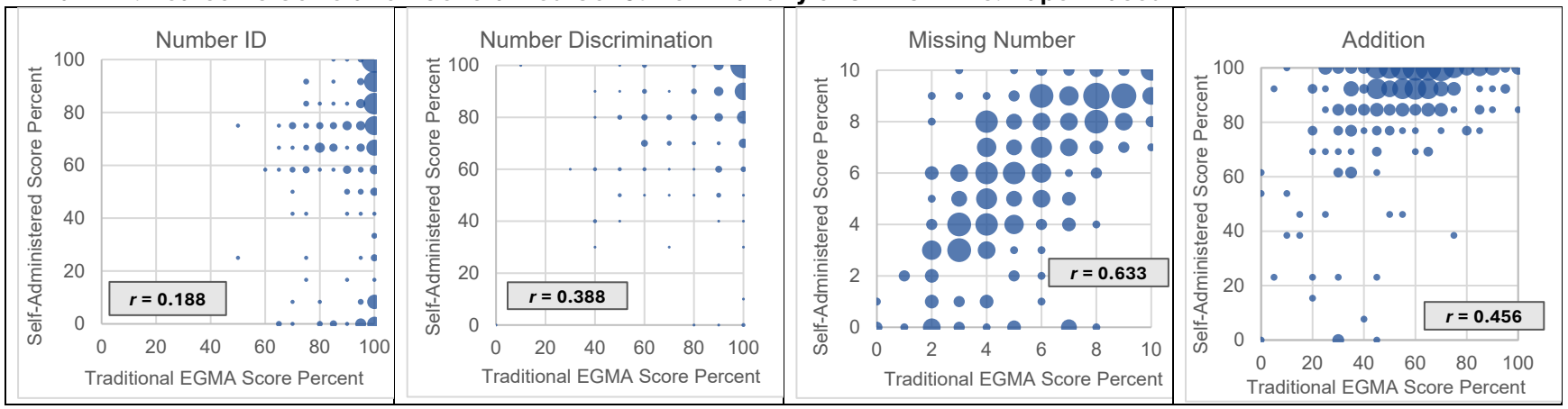


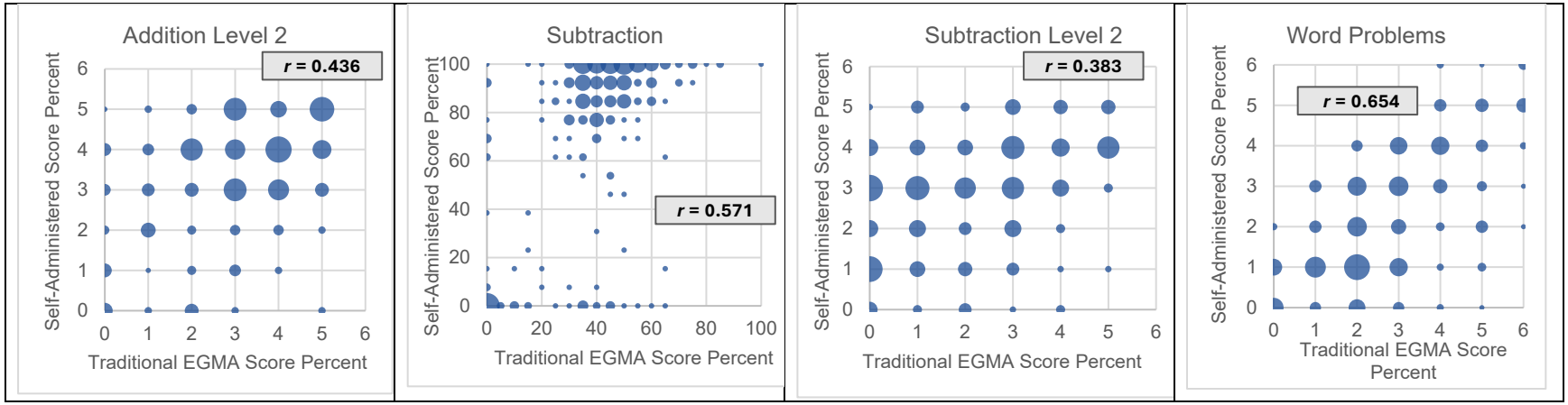
SA-EGMA

Exhibit 0-2: Pearson's Correlation for Generalized Concurrent Validity of the SA-EGMA and Traditional EGMA

SA-EGMA Task Percent Score	Correlation
Number Identification	0.188
Number Discrimination	0.388
Missing Number	0.633
Addition Level 1	0.456
Addition Level 2	0.436
Subtraction Level 1	0.571
Subtraction Level 2	0.383
Word Problems	0.654

Exhibit xxx: Pearson's Correlation Generalized Concurrent Validity of SA-EGMA vs. Paper-Based




Addition Level 2
r = 0.436
Self-Administered Score Percent
Traditional EGMA Score Percent
Subtraction
r = 0.571
Self-Administered Score Percent
Traditional EGMA Score Percent
Subtraction Level 2
r = 0.383
Self-Administered Score Percent
Traditional EGMA Score Percent
Word Problems
r = 0.654
Self-Administered Score Percent
Traditional EGMA Score Percent