Deliverable #4:

Results of Pilot Test in Chichewa

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

August 31, 2023

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

Jennifer Ryan, Karon Harden, Yasmin Sitabkhan, Peggy Dubeck, Elizabeth Marsden, Jori Fafoulas, 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 #2

Executive **Summary**

This report summarizes the findings of an effort to develop and validate tablet-based, self-administered assessments of Chichewa-language foundational literacy and numeracy in the early grades in Malawi. 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 effort to develop these Chichewa instruments took place throughout 2023, beginning with an adaptation workshop with language and curriculum experts in January where items were adapted and revised based on the needed specifications. The tool then went through a round of User Testing with nearly 40 students, focusing on the specific experiences of Malawian students. Minor changes were made to the instrument in response to the feedback from User Testing (and incorporated into previous versions of the tool where applicable), and the revised tools were then Field Tested with over 500 students in each of Grade 2 and Grade 4 at 21 schools in the Zomba District of Malawi. IRT and Factor Analysis from this Field Test led to more minor alterations to the tool and the cutting of some lower performing items. The final version of the tools were then Pilot Tested in June and July to assess internal consistency, reliability, and concurrent validity with the traditional EGRA and EGMA tools. This report presents the results of this final Pilot Testing phase.

The findings are very encouraging, showing high internal consistency across tasks in the tools, and generally acceptable internal consistency within tasks. Both SA-EGRA and SA-EGMA performed well on test-retest reliability, showing students mostly scored consistently across timepoints.

The tools differed from the traditional EGRA and EGMA in that they were not able to assess fluency, but rather focused on accuracy of understanding. Thus, the constructs being measured by the two assessment mediums (tablet-based stimuli vs. paper stimuli) are slightly different, and the correlation between the tools is reduced, but generally acceptable. In the context of Malawi, it is clear that the fluency and automaticity constructs of traditional EGMA Addition and Subtraction tasks differ greatly from the constructs being measured in the corresponding SA-EGMA tasks without a time constraint. With this knowledge, we conclude that the tool is still a valid and acceptable tool, but the tasks should be shortened to reflect their proper use, and it should be clear to those using the SA-EGMA for assessment that these tasks only measure basic numeracy skills and not numeric automaticity as in a traditional EGMA.

We conclude the report with recommendations for small changes or further review and propose to continue to search for new avenues for 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 Chichewa 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 Chichewa SA-EGRA/SA-EGMA tools as well as avenues for further development and refinement for various use cases.

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.

2. Assessment Framework

Understanding the psychometric properties of the SA-EGRA/SA-EGMA tools enables 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. An association of 0.5 or higher is considered strong, and an association between 0.3 and 0.5 is considered acceptable. The second approach is the Bland-Altman[1] analysis, which assesses the level of agreement between pairs of repeated measures across the spectrum of the student ability levels. In this approach, if the level of agreement between measures is consistently less than two standard deviations apart, it is considered strong.

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.[2] We apply factor analysis both within tasks (at the item level) and across tasks (to assess reliability of the overall tools).[3]

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 can measure our constructs of interest despite these differences. We measured concurrent validity between the two assessments using Pearson's *r* and also looking for an association of 0.5 or higher for a strong association or between 0.3 and 0.5 for an acceptable association between the two assessments.

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 earlier document entitled *Draft Assessment Specifications and Prototype (APK) Ready for Field*

[1] 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

[2] 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.

[3] 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

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).[4] An adaptation workshop was held in Zomba, Malawi from January 30-February 3, 2023 with several local linguistic, literacy, and numeracy experts. The experts were gathered with the following workshop objectives:

- Review draft instructions/translations for SA-EGRA and SA-EGMA subtasks.
- Develop context, language, and grade appropriate items for subtasks.
- Develop audio recordings for instructions and test items.

The adaptation workshop was successful, and the resulting subtasks and items that had been adapted according to the SA-EGRA specifications were recorded by Chichewa speakers and imported into the Tangerine tool to be used for testing. 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.

3.2 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 letter sounds aloud from a grid of 100 for up to one minute, on the SA-EGRA this task contains eight items in a receptive, multiple-choice format. For each item, the student is presented with five written letters on the screen and hears the sound associated with one of them. Due to the difficulty in distinguishing some of the Chichewa phonemes in isolation, the prompt the student hears is "Tap the letter for the first sound in …" followed by a consonant-vowel (CV) syllable beginning with the target letter. 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 a broad range of Chichewa phonemes. On the SA-EGRA, this task is untimed and therefore measures only accuracy of recognition, not fluency.

Syllable Recognition

Chichewa phonology is dominated by "open" syllables, that is, syllables that end in a vowel. This structure makes the syllable unit particularly salient and lends itself well to a reading pedagogy focused on decoding syllables and breaking down Chichewa's many long words into their component syllables. Consonant clusters are also extremely common in syllable-onset position (i.e., before the vowel in the syllable), and mastering them is essential to fluent decoding in Chichewa, but consonant clusters cannot be assessed via the simple letter sound task. Undoubtedly for these reasons, the syllable reading task has been a favorite on the traditional Chichewa EGRA for over a decade, and the decision was made to add a syllable recognition task to the SA-EGRA as well. The *Syllable Recognition* task is structured very similarly to the Letter Sound Recognition task; the student is presented with five syllables on the screen, hears one of them, and taps the syllable that s/he hears. In addition to simple syllables (i.e., CV), this task allows for the assessment of common pre-nasalized, labialized, and palatized syllable onsets (e.g.

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

nthi, dwi, nya, respectively). This task has nine items, is untimed, and measures accuracy of recognition.

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 eight words given orally. For the initial prompt, the student hears the word in isolation, hears it used again 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. The words on the spelling task were carefully selected to assess a broad range of common phonemes and syllable structures in Chichewa.

Reading Comprehension

The *Short Story Reading Comprehension* task assesses the student's ability to understand a short text. In the SA-EGRA, the student reads a short story (approximately 60 words long) written at the Grade 2 reading level. The student is instructed to read the story out loud to him/herself. In these respects, the text and the task are similar to those commonly given to assess oral reading fluency (ORF) on the traditional EGRA, though ORF is not measured on the SA-EGRA, due to technological restraints. The student is then presented with six multiple-choice questions about the story one at a time. Each question has four answer options. The questions and answer options (though not the reading passage itself) are presented both in written and oral form. Because the passage is short, the student must answer from memory and look-backs are not allowed. To approximate Grade 2 text readability, the texts from the Malawi Standard 2 Chichewa reader were analyzed for average word and sentence length, two features associated with text difficulty, and the passage on the SA-EGRA was written to similar specifications.

The *Silent Reading Comprehension* task assesses the student's ability to understand a longer text, read silently. The student reads a slightly longer story (approximately 110 words long) written at the Grade 4 reading level. The student is instructed to read silently to him/herself and then answer 10 multiple-choice questions one at a time. Again, each question has four answer options. In order to mitigate the limitations of short-term memory, the student is allowed to look back to the passage at will while answering the questions, though s/he cannot return to previously answered questions. The questions and answer options (though not the reading passage itself) are also presented both in written and oral form. The same method for approximating Grade 4 text readability was applied using the texts from the Malawi Standard 4 Chichewa reader.

Language Proficiency

Finally, the *Vocabulary* and *Syntax* tasks assess two other important aspects of the student's proficiency in Chichewa. 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 don't speak at home.

The *Vocabulary* task captures data about the breadth of the student's receptive knowledge of Chichewa words. The task presents the student with 14 multiple choice items, each with one real word and three pseudoword answer options. 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 Chichewa 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. The target words were “academic” or “text” words that are encountered primarily in text versus in everyday life. Word lists were compiled from the Malawi Standard 1-4 Chichewa readers and analyzed for frequency. A team of Malawi education experts then examined the word lists and chose an initial list of 24 words, eight first encountered in Standard 1 and 2 texts, eight first encountered in Standard 3 texts, and eight first encountered in Standard 4 texts. The team also developed 72 pseudowords with a similar distribution of word lengths to serve as distractors for the 24 target words (three distractors for each real word). The 24 original items were then narrowed down to 14 based on their performance on the pilot test.

The *Syntax* task assesses the student’s knowledge of how words are put together to make meaning in Chichewa. The task presents the student with 10 items with four short sentences each. All of the sentences use basically the same vocabulary but differ primarily in syntax (word order) and morphology (word form). One sentence makes sense (e.g., The child kicked the red ball.) and the others do not (e.g., The ball kicked the red child.). 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 Chichewa syntactic and morphological structures that are essential to parsing Chichewa sentences correctly. A low score on the syntax task indicates that one’s low Chichewa language proficiency is likely a hindrance to reading comprehension rather than (just) poor reading skills.

3.3 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. Most of these tasks are identical to the traditional EGMA, with the only difference being that responses require the child to type their answers instead of speak them aloud to an assessor.

Number Identification

The *Number Identification* (a.k.a. “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”). This task has 12 items of increasing difficulty, and a keypad with numbers 0-9 for the student to use to enter the correct answer When the student has typed in their answer, they press the arrow button to move on to the next item. If the student answers 4 in a row incorrectly, the task ends and they are presented with the next task. This is the same autostop rule as traditional EGMA.

Number Discrimination

The *Number Discrimination* (a.k..a *number size comparison*) task evaluates students' ability to discern between quantities, represented symbolically as numbers, and to identify the largest number (e.g., 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). Each item has three options to choose between, and there are 10 items in the task. If the student answers 4 in a row incorrectly, the task ends and they are presented with the next task.

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 the number line of 0-9 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. If the student answers 4 in a row incorrectly, the task ends and they are presented with the next task.

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. The Level 1 tasks contain 7 items each and Level 2 tasks contain five. Initial Field and Pilot Testing had 13 items each for Addition and Subtraction Level 1, but the subtasks were reduced after analysis, as so many items were not needed to test the construct. All analysis included in this report reflect the reduced subtasks, for the analysis of the full 13 items in these subtasks, refer to ***Annex A***. All *Addition* and *Subtraction* tests are untimed and thus do not assess fluency. The students may use pencil and paper for the Level 2 tests. If the student answers 4 in a row incorrectly, the task ends and they are presented with the next task, is the student scores zero on either Level 1 Addition or Level 1 Subtraction, they do not receive the corresponding Level 2 subtask, similar to traditional EGMA.

Word Problems

The *Word Problems* 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, or manipulatives if available, 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. This task contains six items, If the student answers 4 in a row incorrectly, the task ends and they are presented with the next task.

Grade 4 Math Items

The field testing and pilot testing included a section given to only students in Grade 4 and above.

This section consisted of 6 items assessing more advanced numeracy concepts such as multiplication, division, subtraction of large numbers, fractions, and geometry. Three of the items were open response, asking students to input their response on a number line. The other three items were multiple choice, giving four response options for the student to choose from. Analysis from this subtask is not included in the body of this report, as data analysis showed these items to be too difficult for our current population, and thus added no useful information about student's mathematics skills. A brief description of this subtask and analysis from the Pilot Test can be found in ***Annex B***.

3.3 Stages of Assessment

The development of the tool proceeded in three main stages.

1. **User testing:** this stage occurred during initial and subsequent renderings of the tool. The tool was tested iteratively with small user groups, with feedback collected both via direct observation and explicit questions.
2. **Field testing:** this stage assessed the initial performance of the tool, with an objective to refine the tool and address issues prior to piloting. For the SA-EGRA and SA-EGMA, the field test included analysis of multiple-choice items, assessment duration, protocols, and overall construct validity.
3. **Pilot testing:** in this stage, the tool was fully assessed for its psychometric properties, with a particular focus on reliability and validity. This analysis from the data collected from this assessment form the basis of the main findings detailed in this evaluation report.

User Testing

User testing for this iteration of the tool focused on Malawian students' experience with the interface and assessment, and any needed changes to this specific assessment, as much of the tool had already been refined in previous iterations. The field team in Malawi tested the SA-EGRA and SA-EGMA user interfaces with 38 students from two schools (one urban and one rural school) over a three-day span (March 29-April 3, 2023). The team made minor changes to the user interface (its look, feel, and interactive elements) in response to the user testing feedback. The main changes from this round of feedback were adjustments to the audio functions to keep one audio snippet from playing over another if the student moved on from the screen quickly and increasing the font size to be more legible on smaller tablet screens. These adjustments were also incorporated into previous versions of the tool.

Field Testing

The initial field test was conducted from May 29, 2023 through June 6, 2023 with 580 Grade 2 and 519 Grade 4 students at 21 schools in the Zomba District of Malawi. 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 test 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 C***.

Pilot Testing

The pilot test was conducted from June 27, 2023 through July 7, 2023. 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 3 days after they were initially assessed). Approximately 420 students across Grades 2 to 5 were assessed with the Reading assessments, and approximately 440 students were assessed with the Mathematics assessments, in 16 schools in the Zomba district of Malawi. To assess the tools' performance across all ability ranges, we sought an even spread (i.e., a uniform distribution) of student abilities in the reading assessment, and the fewest zero scores possible. 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. Because of this, most students assessed were from Grades 3 and 5, although the tool was still created to align with Malawi Grade 2 and 4 curricula.

4. Pilot Test Findings

4.1 Student literacy outcomes

The student literacy average percent scores should be viewed in context of the sampling approach used to select students. The students were selected purposefully each day to develop an approximately uniform distribution of student performance along an oral reading fluency scale. Consequently, the average percentage 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.

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

SA-EGRA Task	Average Percent Score
Syntax	74.9%
Letter Sounds	80.5%
Vocabulary	66.7%
Spelling	52.5%
Silent Reading Comprehension	50.2%
Short Story Reading Comprehension	68.6%
Syllables	68.6%

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; 74.9%) and Letter Sounds (80.5%). The most challenging tasks were Spelling (the only productive / non-multiple-choice task; 52.5%) and the higher-order literacy skill of comprehension (50.2%).

4.2 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 and 5 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, with scores decreasing as the difficulty of subtasks increases.

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 of the Ghanaian assessment, where student responses were indicative of students learning the input features of the tablet, including by entering long sequences of random multiple-digit numbers. 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 in previous iterations to alleviate this effect, which helped minimize the number of occurrences, but it is still not wholly eliminated. We are heartened to see that this behavior is less between timepoint one and timepoint two of the test-retest, so it may be possible that students who are more exposed to the assessment will be less likely to repeat these patterns. This behavior will need to be considered in future uses of the tool for scoring purposes.

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

SA-EGMA Task	Average Percent Score
Number Identification	77.4%
Number Discrimination	84.3%
Missing Number	53.6%
Addition Level 1	82.7%
Addition Level 2	61.8%
Subtraction Level 1	76.9%
Subtraction Level 2	44.5%
Word Problems	51.2%

4.3 Time taken to complete assessment

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.

Exhibit 3: Mean Assessment Durations for SA-EGRA and SA-EGMA, by time point

Assessment	Mean (mins.)	Standard Deviation
SA-EGRA (t_1)	42.4	±0.70
SA-EGRA (t_2)	35.7	±0.71
SA-EGMA (t_1)	38.3	±0.83
SA-EGMA (t_2)	31.7	±0.80

The mean duration of students' SA-EGRA and SA-EGMA assessments both reduced a fair amount between timepoints. This points to a possible learning curve in using the assessment for these students. The RTI team recommends shortening the assessment when possible, to prevent test fatigue, and including additional training for supervising adults to instruct students in methods to move on if they are stuck on any one item or task. One recommendation we have implemented following this analysis is to include autostop rules for the SA-EGRA that will help students who struggle to move through the assessment quicker. These recommendations will be discussed in the recommendations section below.

4.4 Internal Consistency

SA-EGRA Tasks

Here we present a summary of the tools' internal consistency; the detailed analyses are provided in ***Annex D***. 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*** below, found strong factor loadings into a single construct ranging from 0.7032 (*Letter Sounds*) to 0.8402 (*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.

Exhibit 4: Factor Analysis Loadings for SA-EGRA task scores

SA-EGRA Task Percent Score	Factor 1 Loadings
Syntax	0.714
Letter Sounds	0.703
Vocabulary	0.715
Spelling	0.840
Silent Reading Comprehension	0.737
Short Story Reading Comprehension	0.760
Syllables	0.775

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 D***. The discussion here will be limited to key points in summary form.

For IRT to work well, the items being analyzed should explain the same construct. In this case, item 15 has a factor loading less than 0.3, suggesting that this item either explains a different construct or may need adjustment. It is not immediately clear why this item has a lower factor loading than others, aside from also being a higher difficulty and higher discrimination item. For some reason, this item is easy for high performers, but very difficult for lower performers and does not fit as well into the construct of the subtask. It is possible that it is an effect from testing fatigue, being the last item in the subtask, but we recommend specific review of this item by Chichewa literacy experts to understand why some students may find it so easy while others struggle so much.

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

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	*Difficulty*	*Bi-serial Correlation*
1	0.581	0.71	0.72	0.65
3	0.685	0.73	0.74	0.72
5	0.640	0.55	0.81	0.67
6	0.653	0.48	0.85	0.68
8	0.484	0.73	0.67	0.58
10	0.661	0.56	0.82	0.69
12	0.727	0.73	0.77	0.74
13	0.609	0.65	0.77	0.66
14	0.679	0.66	0.78	0.70
15	0.279	0.70	0.56	0.43

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 0.48 (Item 6) to 0.73 (Items 8 and 12). 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. Most of the *Syntax* items fall within this range, with the remainder of the items exhibiting *Difficulty* scores of over 0.8. *Syntax* was the second highest-scoring task behind *Letter Sounds*, so these *Difficulty* scores should be interpreted in the context of this being a generally easier subtask than others. 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 0.43 to 0.74. Taken together, these analyses point to an acceptably strong subtask that can be used to assess lower-level literacy skills in students, while also leaving room for iterative improvement of certain items such as item 15 as noted above.

Internal consistency for *Letter Sounds* was acceptable overall (***Exhibit D3 of Annex D***). All factor loadings were in acceptable range, with each item having appropriate levels of *Discrimination* and *Difficulty*, although there was not a wide range across these measures. This lack of variability in *Discrimination* and *Difficulty* is most likely due to the task targeting lower-order literacy skills, and our sample population needing to consist of mostly Grade 3 and 5 students.

The internal consistency for the *Vocabulary* task is shown in ***Exhibit D4*** of ***Annex D***. The *Discrimination* and *Difficulty* scores are acceptable, with only item 17 having a *Difficulty* over 0.8 (0.85). All factor loadings fall into the acceptable range for good internal consistency.

For *Silent Reading Comprehension*, while two items (9, and 11) display low factor loading and merit further review, *Discrimination* and *Difficulty* are within acceptable ranges for all items, with only one (Item 5) having a *Difficulty* score over 0.8.

The *Syllables* task also shows good internal consistency, with only item 6 displaying a low factor loading. This item stands out as the hardest Syllables item with only 55% of students chose the correct answer, compared to a range of 64% to 79% correct for other items. Item 6 should be reviewed by a Chichewa language expert to ensure there are no issues with the audio or how the item was created that may create difficulty for students to understand the correct answer.

For *Short Story Reading Comprehension*, while question 6 displays a borderline low factor loading and may 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. Three-parameter 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 D8*** of ***Annex D***) is excellent, with factor loads being between 0.814 and 0.870, 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 96 (a wide continuous measure). Consequently, the *Spelling* task is excellent for capturing the variability of students' skill levels.

Overall, the internal consistency of the SA-EGRA tasks are strong, and the analysis all points to cohesive subtasks that are measuring the same general construct of literacy. This, combined with the internal consistency of the full assessment, point to a strong instrument for assessing literacy in Chichewa-speaking Grade 2-4 students in Malawi.

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 0.5089 (*Word Problems*) to 0.7575 (*Addition*). Given that acceptable factor scores should be 0.3 or more, the internal consistency of the tasks for the SA-EGMA is good and the tasks are all performing well together to measure numeracy.

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

Task Percent Score	Factor 1 Loadings
Number Identification	0.646
Number Discrimination	0.656
Missing Number	0.734
Addition Level 1	0.681
Addition Level 2	0.733
Subtraction Level 1	0.503
Subtraction Level 2	0.639
Word Problems	0.504

IRT techniques are a poor fit for free-response item types. With the exception of *Number Discrimination*, all SA-EGMA tasks 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 *Addition Levels 1 and 2* tasks

Item	Factor Analysis
1	0.248
2	0.272
3	0.049
4	0.139
5	0.151
6	0.460
7	0.374
8	0.331
9	0.269
10	0.403
11	0.496
12	0.541

Exhibit 7 at left presents the item-level factor analysis for the *Addition Level 1* 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 12 items surpass the target threshold (a loading of 0.3 on the first factor). Item number 3 (i.e., 3 + 3), has a worryingly low loading, but further analysis reveals nothing out of the ordinary for this item, other than being the first item for which low performing learners could be auto stopped, as in the full assessment, it was the fifth item. As discussed above, the auto stop rule for this task is 4 incorrect items in a row, and 5.6% of students were stopped before item 5.

Item-level factor 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 -0.075). The factor loadings for individual items within a subtask are not very high, although this is common in Traditional EGMA assessments and to be expected. While the subtasks are all measuring general foundational skills, each subtasks measures distinct skills, and within subtasks, items are designed to measure easier to harder tasks. This progression within subtasks means that the items within a subtask are not designed to measure the same skill, but rather a progression of a particular skill. For example, within quantity discrimination, each task within the subtask is increasingly more difficult, moving from simple quantity discrimination of single-digit numbers, to double-digit numbers, and then triple-digit numbers, with different specifications for the types of numbers used in the task according to developmental progressions. Given this, lower factor loadings within a subtask are not cause for concern but can still provide information on where improvements may be made to a task.

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. The item with the lowest factor loading in the *Word Problem* task, item 1, most likely has a lower loading because this item was easier for students than the other items. This is still an important item to include to ensure students understand the task and what they are asked to do.

4.5Test-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. An acceptable range for the Pearson's coefficient would be between ±0.3 and ±1, with 0.3 to 0.49 indicating a moderately strong relationship and 0.5 to 1 indicating a high correlation. ***Exhibit 8*** at the 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 E, Exhibit E2***.

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

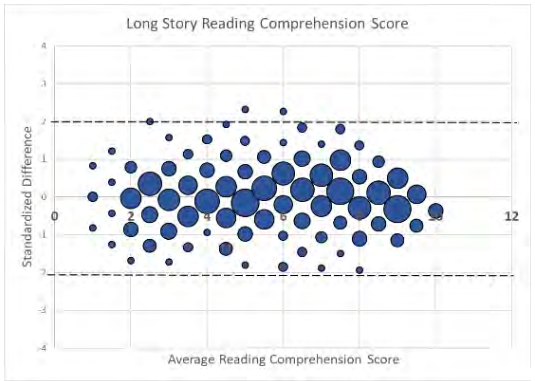
SA-EGRA Task Percent Score	Correlation
Letter Sounds	0.795
Short Reading Comprehension	0.729
Syllables	0.771
Syntax	0.703
Long Reading Comprehension	0.742
Vocabulary	0.773
Spelling	0.919

The *Spelling* task demonstrates the strongest positive correlation at the two timepoints. 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 **D6**). The other tasks demonstrate correlations in the range of approximately 0.7 to 0.8.

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 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 this principle, ***Exhibit 9*** below presents the Bland-Altman graph for students' scores on the *Long Story Reading Comprehension* 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.[5] 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 *Long Story Reading Comprehension*. The Bland-Altman graphs for the other tasks are included in ***Annex E*.**

Exhibit 9: Bland-Altman Graph for SA-EGRA Test-Retest,
***Long Story Reading Comprehension* task**



The Bland-Altman plots and Pearson correlation plots indicate that the SA-EGRA tasks demonstrate very good test-retest reliability (Annex C). Notably, only a small percentage of students scored very well at one timepoint of a subtask 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. As this phenomenon only accounts for less than 4% of students at the most, we do not believe it will cause issues with data collection or overall inaccuracies with the assessment but may be an opportunity for future improvements to the tool. 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** on the 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**.

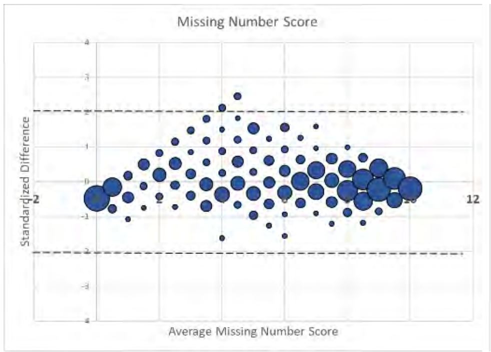
Exhibit 9: Pearson's Correlation for the SA-EGMA Test-Retest

SA-EGRA Task Percent Score	Correlation
Number Identification	0.676
Number Discrimination	0.724
Missing Number	0.824
Addition Level 1	0.697
Addition Level 2	0.612
Subtraction Level 1	0.630
Subtraction Level 2	0.474
Word Problems	0.676

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 Pearson's correlations; we again complement the analysis with Bland-Altman plots. The plot for *Missing Number*, 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.

[5] 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 10: Bland-Altman Graph for SA-EGMA Test-Retest, *Missing Number* task



Most of the students fall within the expected range of ±2 standard deviations, with less than 1% falling outside the range. The larger points, indicating a larger percentage of students, cluster around 0 on the vertical axis, particularly at both ends of the horizontal axis, indicating that most students scored very similarly at both timepoints. The pattern of small points fanning upwards shows that a portion of students scored low at one timepoint and high at another (although not outside the range of tolerance). The Pearson's Correlation graph in Exhibit C-5 shows that a small number of students scored poorly in the first timepoint and well in the second timepoint, possibly indicating an unfamiliarity with the task initially. Prior to using this task to assess numeracy in children, it would be prudent to review how familiar students may be with number patterns and possibly add another practice item in areas where children are not as familiar with this sort of task. Aside from this consideration, the Missing Number task performs very well and can be used as a good metric for assessing number sense in students.

The Test-Retest Reliability of the Self-Administered tasks show that all are within the acceptable range, and each task is a reliable measure of the literacy or numeracy construct they are measuring, with very few caveats.

4.6 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 (on 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.[6]

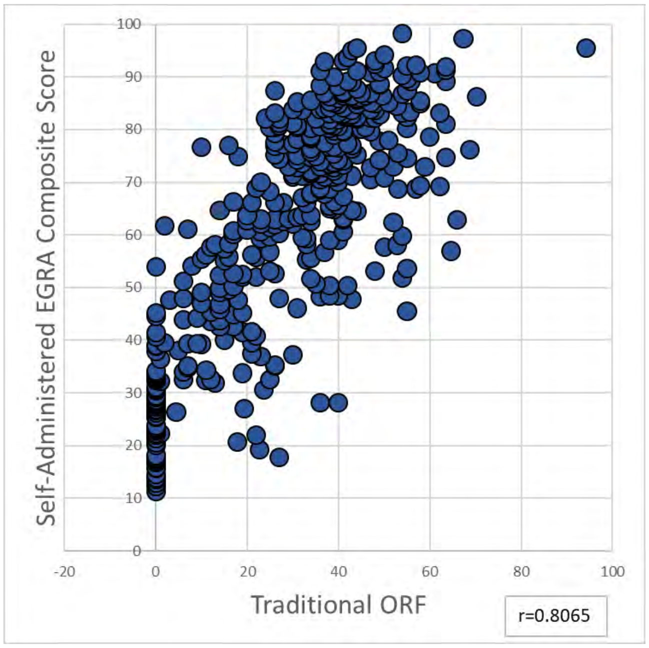
We therefore sought a method of exploring SA-EGRA's concurrent validity with the traditional EGRA that would be likely to meet the needs of the SA-EGRA's target user base. We decided to explore whether the SA-EGRA can be used as a proxy measure for ORF. One option would be to create a composite SA-EGRA score, combining the task percent scores weighted according to

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

relative importance derived from expert judgment.

The correlation scatterplot comparing the scores on the SA-EGRA Composite Score and the traditional EGRA's ORF task is shown in ***Exhibit 12***, below. The SA-EGRA Composite Score was calculated using this formula: SA_EGRA_composite = 0.4*spelling_total_score_pcnt+0.15*short_read_comp_score_pcnt+0.1*long_read_comp_score_pcnt+0.05*letter_sounds_score_pcnt+0.1*vocab_score_pcnt+0.1*syntax_score_pcnt+0.1*syllables_score_pcnt

Exhibit 11 SA-EGRA Composite Score vs. ORF on the Traditional EGRA



The correlation of *r* = 0.8065 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 a 60 (rounded, 59.5-60.4) Composite Score recorded ORF scores between 71 and 54 correct words per minute. This suggests that while this statistical linking of SA-EGRA *Composite Score* 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 traditional EGRA for individual students. Based on the excellent performance of the *Spelling* task, we also elected to assess how well it is associated with ORF on the traditional EGRA, and those results can be seen in Annex D. The statistical linking to this task ended in similar conclusions, that it is good for generalized results of a large group, but not for individual students, although the positive linear association was slightly weaker. Having performed this analysis across two contexts, we have determined that the SA-EGRA *Composite Score* may be the better option for using as a single metric for literacy from this assessment. The formula used to create the composite score was used again on previous data from Pilot Testing in Ghana and also performed very well, even though that assessment did not include a syllables subtask. We therefore recommend using the SA-EGRA *Composite Score* as an appropriate measure for literacy and mapping to traditional EGRA scores for population estimates.

SA-EGMA Tasks

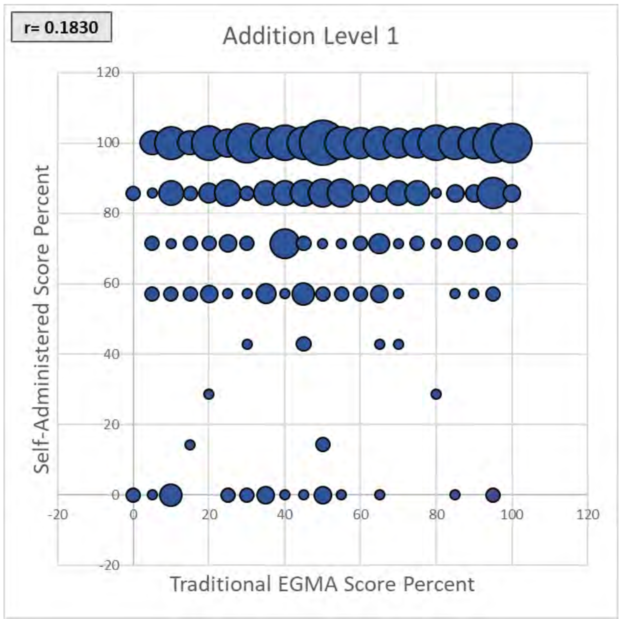
There is a very tight coupling at both the task and item levels between the SA-EGMA and the traditional EGMA. As a result, using Pearson's correlation for generalized assessment of concurrent validity is a more appropriate method than was the case for the SA-EGRA and traditional EGRA. ***Exhibit 13*** above presents the Pearson's correlation for each task. As with the test-retest correlations, an acceptable range for the Pearson's coefficient would be between ±0.3 and ±1, with 0.3 to 0.49 indicating a moderate correlation and 0.5 to 1 indicating a strong relationship.

Exhibit 12: Pearson's Correlation for Generalized Concurrent Validity of the SA-EGMA and Traditional EGMA

SA-EGMA Task Percent Score	Correlation
Number Identification	0.511
Number Discrimination	0.560
Missing Number	0.676
Addition Level 1	0.183
Addition Level 2	0.467
Subtraction Level 1	0.203
Subtraction Level 2	0.284
Word Problems	0.581

The most notable feature of the SA-EGMA's task-level correlations with the traditional EGMA is how widely they range. The *Addition Level 1* is very weakly correlated (*r* = 0.183), while *Missing Number* (*r* = 0.676) and *Word Problems* (*r* = 0.581) are much more strongly correlated. Given how little the items themselves differed across assessments, this suggests that the differences in administration modality have a substantial influence over students' math scores. It seems that having an assessor in front of the student, timing them and nudging them onwards through the task, in addition to timing the student for fluency, has a great impact on student performance on the items. This influence is seen very clearly in the scatterplot of the correlation between traditional EGMA *Addition Level 1* scores and SA-EGMA *Addition Level 1* scores in **Exhibit 14** below.

Exhibit 14: Pearson's Correlation Graph for Concurrent Validity, *Addition Level 1* task



There is a ceiling of scores across the top, showing a large percentage of students scoring highly on the self-administered task, but performing across the range of scores in the traditional task. This shows that, for a large portion of students, self-administered *Addition Level 1* scores do not accurately reflect traditional EGMA scores, and students tend to score higher on the self-administered than the traditional task. We believe this is due to the difference in underlying constructs created by the removal of time limits from the Self-Administered EGMA tasks. In a typical EGMA, the *Addition Level 1* task is designed to measure fluency and automaticity by asking the student to complete as many items as possible in 60 seconds. Because it is too difficult to disentangle the time it takes a child to solve an addition task and the time it takes for them to record their answer and move on to the next task, we cannot sufficiently measure these constructs in the self-administered EGMA and are therefore only measuring knowledge of basic mathematics skills. This disconnect between the constructs measured by these two assessments was not as pronounced in previous iterations of the SA-EGMA, and we believe that it is more apparent in contexts where children have lower levels of numerical fluency and automaticity but have still mastered some basic mathematics skills. We do not recommend using the Addition or Subtraction tasks to predict traditional EGMA scores because of these differences. Because this slightly lower concurrent validity has been seen on more than one iteration of the self-administered EGMA, and because we can now see the different constructs being measured by the two tools in this context, we also recommend reducing the number of items in the *Addition Level 1* and *Subtraction Level 1* tasks to the 7 items we have reviewed in this report, as more are not necessary for measuring knowledge of these skills without the 60-second time limit used in the traditional assessment. We can reliably assess a student's basic numeracy skills with fewer items in the task, allowing for a shorter assessment. We still believe the SA-EGMA Addition and Subtraction tasks are valid measures of numeracy and can still be used for assessment, but we no longer need so many items.

While other tasks were less successful in replication, we can recommend using the Missing Number subtask and Word Problems subtask as standalone subtasks for assessing student's emerging numeracy and number sense, and we can recommend the SA-EGMA tool as a valid and reliable tool for assessing numeracy skills in a population.

5 Conclusions and Recommendations

The primary aim of the effort that culminated in this report was to develop *tablet-based assessments* 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:

The internal consistency of the SA-EGRA as an assessment, and within each of the individual tasks, is strong and points to consistent tasks and a cohesive assessment of literacy. We recommend that Syllables Item 6 be reviewed by a Chichewa literacy expert to ensure that the task is performing appropriately and there are no issues with the audio causing confusion for students. Analysis of Test-Retest Reliability shows that the SA-EGRA is also a reliable measure of Chichewa literacy in Grades 2-4 in Malawi. The SA-EGRA also shows strong construct validity, when the Composite Scores are correlated with traditional ORF scores, even though it cannot predict them at the individual level. The scores from this instrument can be used to create generalized traditional ORF estimates at the population level but should not be used to predict scores for individual students. This Chichewa SA-EGRA is a reliable, valid tool for assessing Early Grade Literacy in Grade 2-4 students in Malawi.

Recommendations:

- When it is possible, shorten the assessment to only the needed subtasks for the grade and curriculum to be assessed, to prevent test fatigue.
- Implement autostop rules to have students who cannot score higher than 3 correct letters on *Letter Sounds* or *Syllables* skip the Reading Comprehension passages and receive a zero score for reading.
- Train supervising adults to instruct students in methods to move on if they are stuck on any one item or task.

- Syllables Item 6 should be reviewed by a Chichewa language expert to ensure there are no issues with the audio or how the item was created that may create difficulty for students to understand the correct answer.

Mathematics:
The SA-EGMA was found to be internally consistent as an assessment with the factor analysis of the task percent scores showing high factor loadings on a single factor for all tasks. However, within the tasks, many items had factor loadings lower than ideal, but are not outside the realm of Traditional EGMA results. Test-retest analysis found the SA-EGMA to be a reliable assessment across timepoints, with students scoring very similarly. The validity of the SA-EGMA was generally very good, with most tasks correlating highly with traditional EGMA tasks assessing the same constructs. The concurrent validity analysis made clear that the traditional EGMA and the SA-EGMA are measuring very different constructs in the Addition and Subtraction tasks without the ability to time students to measure fluency and automaticity in the self-administered version. Because of this, we do not recommend using the SA-EGMA Addition and Subtraction tasks to predict traditional EGMA scores, but we still believe the SA-EGMA tool is a valid method to assess basic numeracy skills and recommend Missing Number and Word Problems as valid and reliable measures of number sense and basic numeracy skills.

Recommendations:

- When it is possible, shorten the assessment to only the needed subtasks for the grade and curriculum to be assessed, to prevent test fatigue.
- Anyone using the assessment should review how familiar students who they wish to assess are with number patterns and tasks similar to the Missing Number subtask and add in extra practice items in areas where children are less familiar with this kind of numeracy task.
- Reduce the number of items in Addition Level 1 and Subtraction Level 1 to the 7 items tested in this report, as we are no longer attempting to measure fluency or automaticity, but simply basic numeracy skills. (This recommendation has already been implemented).

The SA-EGRA and SA-EGRA overall performance was very encouraging, and we believe the evidence supports a conclusion that it is appropriate to deploy the tools with the minor modifications already implemented for assessing Chichewa literacy in early grades in Malawi. That said, we recommend being open to further opportunities to iteratively revise these tools as evidence accumulates about their performance in various contexts.

Annex A. Addition and Subtraction Level 1 Analysis

This Annex reviews the full suite of items in the initial Addition Level 1 and Subtraction Level 1 subtasks. Items 2, 4, 6, 9, 10, and 13 were cut from each, to leave each task with only 7 items. As the purpose of the subtask has changed to no longer measure fluency, multiple items with identical or similar specifications (measuring the same sub-skill) do not need to be included. These items were chosen to drop because they were targeted the same sub-skills as other items.

Exhibit A1: Item Factor Analysis for the *Addition Levels 1 and 2* tasks

Item	Label	Factor Analysis
1	1 + 3 = (4)	0.234
2	2 + 3 = (5)	0.117
3	6 + 2 = (8)	0.332
4	4 + 5 = (9)	0.254
5	3 + 3 = (6)	0.070
6	7 + 3 = (10)	0.318
7	8 + 1 = (9)	0.156
8	2 + 8 = (10)	0.314
9	7 + 5 = (12)	0.381
10	8 + 6 = (14)	0.299
11	9 + 8 = (17)	0.425
12	10 + 2 = (12)	0.300
13	8 + 10 = (18)	0.529
14	13 + 6 = (19)	0.231
15	18 + 7 = (25)	0.254
16	12 + 14 = (26)	0.421
17	22 + 37 = (59)	0.450
18	38 + 26 = (64)	0.497

Exhibit A2: Item Factor Analysis for the *Subtraction Levels 1 and 2* tasks

Item	Label	Factor Analysis
1	4 – 3 = (1)	0.13
2	5 – 3 = (2)	-0.0526
3	8 – 2 = (6)	0.334
4	9 – 5 = (4)	0.2339
5	6 – 3 = (3)	0.0955
6	10 – 3 = (7)	0.1701
7	9 – 1= (8)	0.1971
8	10 – 8 = (2)	0.0459
9	12 – 5 = (7)	0.1436
10	14 – 6 = (8)	0.3695

11	17 – 8 = (9)	0.3687
12	12 – 2 = (10)	0.3057
13	18 – 10 = (8)	0.3424
14	19 – 6 = (13)	0.2048
15	25 – 7 = (18)	0.1816
16	26 – 14 = (12)	0.3875
17	59 – 37 = (22)	0.3624
18	64 – 26 = (38)	0.272

Exhibit A3: Pearson's Correlation Generalized Test-Retest Reliability for the *Addition and Subtraction Level 1* tasks

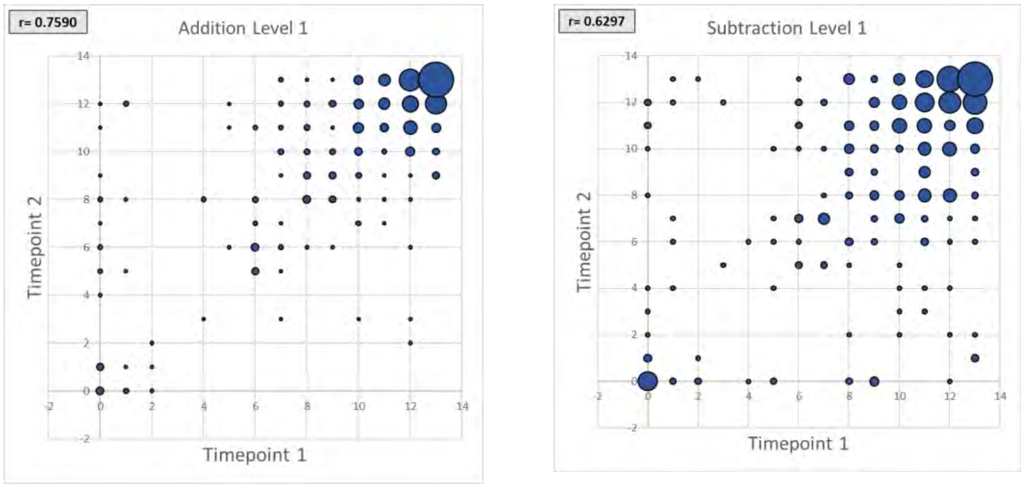
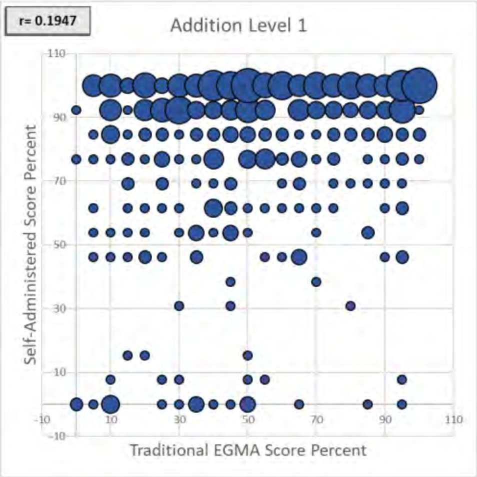
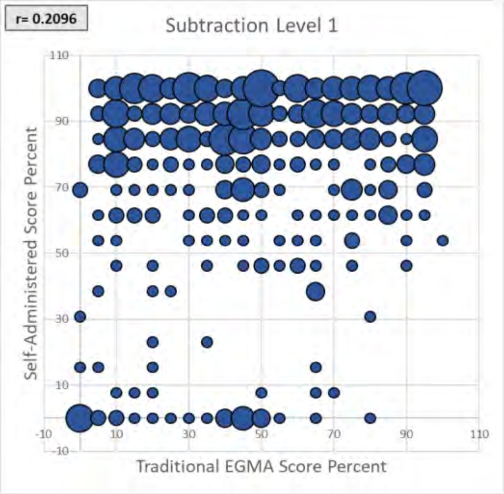


Exhibit A4: Pearson's Correlation Generalized Concurrent Validity of SA-EGMA vs. Paper-Based for *Addition and Subtraction Level 1* tasks

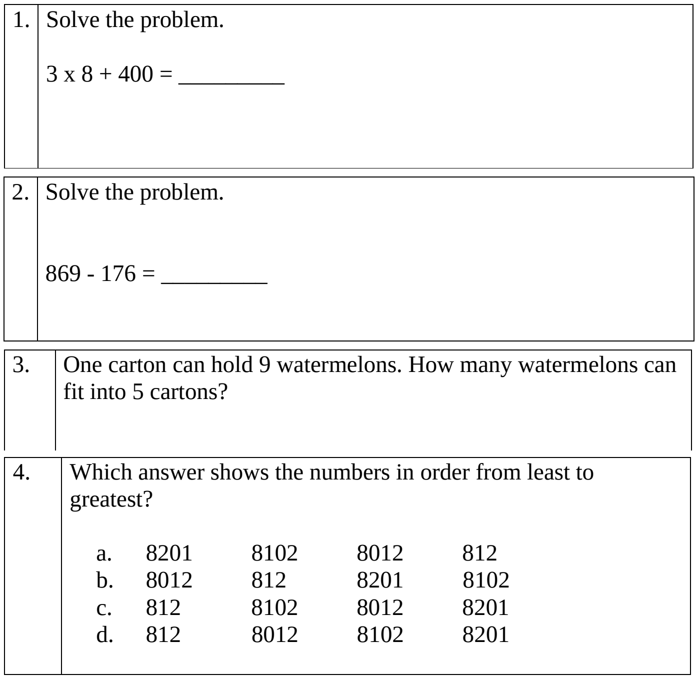

r= 0.1947
Addition Level 1
Self-Administered Score Percent
Traditional EGMA Score Percent
110
90
70
50
30
10
-10
-10
10
30
50
70
90
110


r= 0.2096
Subtraction Level 1
Self-Administered Score Percent
Traditional EGMA Score Percent
110
90
70
50
30
10
-10
-10
10
30
50
70
90
110

Annex B. SA-EGMA Grade 4 Math Subtask and Analysis

The Grade 4 Math Subtask was only given to students in Grades 4 and above and consisted of 6 items assessing more advanced numeracy concepts such as multiplication, division, subtaction of large numbers, fractions, and geometry. Three of the items were open response, asking students to input their response on a number line. The other three items were multiple choice, giving four response options for the student to choose from. The specific items can be seen below in Exhibit B.1.

Exhibit B1: SA-EGMA Grade 4 Math Items



1. Solve the problem.

3 x 8 + 400 = _________

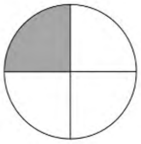
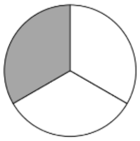
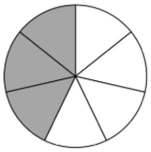
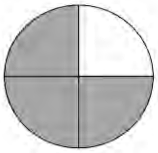
2. Solve the problem.

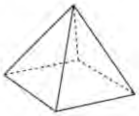
869 - 176 = _________

3. One carton can hold 9 watermelons. How many watermelons can fit into 5 cartons?

4. Which answer shows the numbers in order from least to greatest?

a. 8201 8102 8012 812
b. 8012 812 8201 8102
c. 812 8102 8012 8201
d. 812 8012 8102 8201

5.	Which figure shows three quarters of a circle shaded?
a.	
b.	
c.	
d.	

6.	What is the figure shown below called? __________ 	
	a	triangle
	b	cone
	c	pyramid
	d	cylinder

These items were included to capture advanced mathematics skills of students in older grades, but in this context, they were inappropriate for this purpose. Students were not able to answer any of the open response items correctly, and the multiple-choice items showed students struggled with the difficulty of the items, also. Item 4 had only slightly above 25% of students answer correctly, which is the rate we would expect for students randomly choosing an item. Item 6 had even less students than random choice answer correctly, as most students chose the distractor item of "triangle". This shows that Grade 4 and 5 students in this context know two-dimensional shapes, but have not yet been taught 3-dimensional, giving further evidence that this task is too difficult for these students. Students scored an average of 14.3% correct on this subtask.

Exhibit B2: SA-EGMA Grade 4 Math Item Scores

Grade 4 Math Item	Average Percent Score
Item 1	0%
Item 2	0%
Item 3	0%
Item 4	26.4%
Item 5	45.2%

Item 6	14.4%
Average Overall Score	14.3%

The Grade 4 Math Subtask did not perform well in any Factor Analysis or Cronbach's Alpha Analysis of the SA-EGMA and thus was discarded from the tool.

Test-Retest analysis found the task to perform poorly as well. with a Pearson's Correlation of 0.27, and a mediocre level of agreement on the Bland-Altman analyses. The graphs of the Pearson's Correlation analysis and Bland-Altman analysis are below in

Exhibit B3: Pearson's Correlation Generalized Test-Retest Reliability for Grade 4 Math Subtask

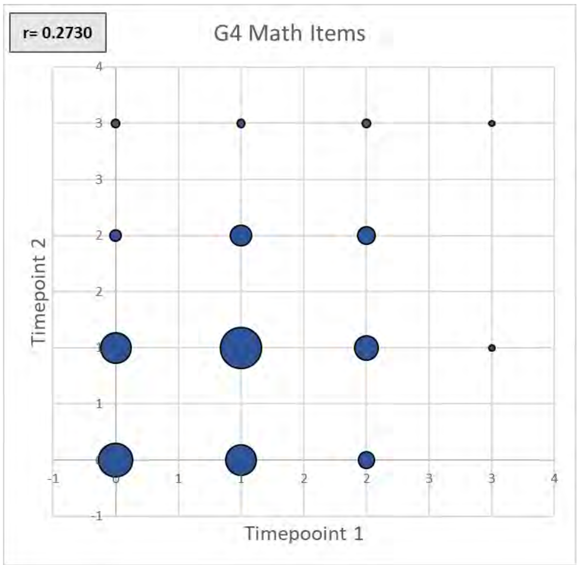
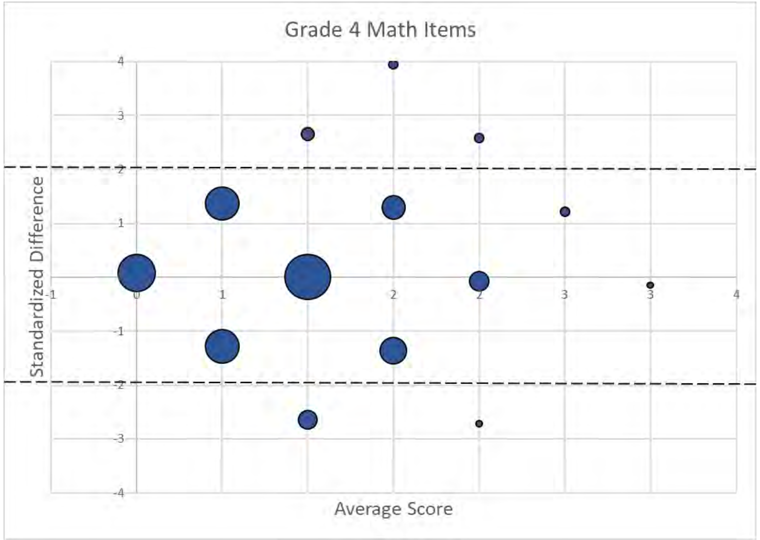


Exhibit B4: Bland-Altman Plots of Test-Retest Reliability for Grade 4 Math Subtask



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

Exhibit C1: Field Test Recommendations for modifications to the SA-EGRA

Task	Recommendations
Letter Sounds	Remove two items.
Short Story Reading Comprehension	Incorporate autostop rules.
Syllables	Remove one item.
Spelling	Remove four items.
Silent Reading Comprehension	Remove one item, switch position of two items. Incorporate autostop rules.
Vocabulary	Remove ten items.
Syntax	Remove five items.

Exhibit C2: Field Test Recommendations for modifications to the SA-EGMA

Task	Recommendations
Number Identification	No change.
Number Discrimination	No change.
Missing Number	No change.
Addition	Remove six items.
Addition Level 2	No change.
Subtraction	Remove six items.
Subtraction Level 2	No change.
Word Problems	No change.

Annex D. 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 0.70 (*Letter Sounds* task percent score) to 0.84 (*Spelling* task percent score). The task-level internal consistency of the SA-EGRA is excellent.

Exhibit D1: Factor Analysis Loadings for SA-EGRA task scores

SA-EGRA Task Percent Score	Factor 1 Loadings
Letter Sounds	0.7032
Short Story Reading Comprehension	0.76
Syllables	0.7754
Spelling	0.8402
Silent Reading Comprehension	0.7372
Vocabulary	0.7145
Syntax	0.7136

Exhibit D13: Item Factor Analysis and IRT for the *Syntax* task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
1	0.408	0.71	0.72	0.65
3	0.534	0.73	0.74	0.72
5	0.474	0.55	0.81	0.67
6	0.496	0.48	0.85	0.68
8	0.299	0.73	0.67	0.58
10	0.506	0.56	0.82	0.69
12	0.576	0.73	0.77	0.74
13	0.436	0.65	0.77	0.66
14	0.516	0.66	0.78	0.7
15	0.108	0.7	0.56	0.43

Exhibit D14: Item Factor Analysis and IRT for the *Letter Sounds* task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
1	0.379	0.5	0.8	0.51
3	0.603	0.51	0.88	0.63
4	0.413	0.75	0.63	0.58
5	0.614	0.63	0.81	0.67
6	0.485	0.67	0.75	0.59
7	0.507	0.51	0.85	0.59
9	0.584	0.46	0.88	0.63
10	0.503	0.41	0.88	0.58

Exhibit Error! No text of specified style in document.**D15: Item Factor Analysis and IRT for the *Vocabulary* task**

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
1	0.551	0.71	0.68	0.6
6	0.331	0.61	0.39	0.43
7	0.503	0.54	0.79	0.53
9	0.583	0.8	0.63	0.62
10	0.532	0.6	0.75	0.56
13	0.394	0.69	0.57	0.48
16	0.462	0.78	0.53	0.54
17	0.508	0.47	0.85	0.52
18	0.565	0.75	0.71	0.6
19	0.438	0.52	0.78	0.49
20	0.545	0.7	0.64	0.6
21	0.510	0.82	0.51	0.58
22	0.581	0.62	0.74	0.61
24	0.623	0.67	0.76	0.65

Exhibit D5*:* Item Factor Analysis and IRT for the *Silent Reading Comprehension* task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
1	0.475	0.7	0.58	0.55
2	0.535	0.72	0.72	0.58
3	0.503	0.72	0.73	0.55
4	0.441	0.74	0.57	0.53
5	0.404	0.49	0.81	0.45

6	0.360	0.65	0.41	0.48
7	0.467	0.73	0.57	0.55
8	0.382	0.7	0.37	0.5
9	0.196	0.49	0.48	0.37
11	0.189	0.4	0.28	0.36

Exhibit D6: Item Factor Analysis and IRT for the *Syllables* task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
1	0.484	0.49	0.79	0.56
2	0.521	0.62	0.64	0.58
3	0.456	0.57	0.72	0.55
4	0.505	0.54	0.76	0.58
5	0.402	0.63	0.55	0.53
6	0.157	0.46	0.42	0.36
7	0.436	0.44	0.75	0.52
9	0.370	0.45	0.74	0.49
10	0.617	0.55	0.8	0.65

Exhibit D7: Item Factor Analysis and IRT for the *Short Story Reading Comprehension* task

Item Number	Factor Analysis	Item Response Theory		
		Discrimination	Difficulty	Bi-serial Correlation
1	0.438	0.71	0.74	0.58
2	0.526	0.76	0.67	0.64
3	0.540	0.87	0.58	0.66
4	0.447	0.69	0.77	0.58
5	0.568	0.87	0.58	0.68
6	0.293	0.55	0.79	0.46

Exhibit D8: Item Factor Analysis for the *Spelling* task

Item	Factor Analysis
2	0.814
4	0.853
6	0.869
8	0.834
9	0.866
10	0.870

11	0.834
12	0.835

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 0.504 (*Word Problems* task percent score) to 0.734 (*Missing Number* task percent score). The task-level internal consistency of the SA-EGMA is good, albeit less strong than the SA-EGRA.

Exhibit D9: Factor Analysis Loadings for SA-EGMA task scores

Task Percent Score	Factor 1 Loadings
Number Identification	0.646
Number Discrimination	0.656
Missing Number	0.734
Addition	0.681
Addition Level 2	0.733
Subtraction	0.503
Subtraction Level 2	0.639
Word Problems	0.504

Exhibit D10: Item Factor Analysis for the *Number Identification* task

Item	Factor Analysis
1	0.237
2	0.289
3	0.380
4	0.230
5	0.142
6	0.198
7	0.364
8	0.224
9	0.350
10	0.430
11	0.600
12	0.572

Exhibit D11: Item Factor Analysis for the *Number Discrimination* task

Item	Factor Analysis
1	0.139
2	0.277
3	0.174
4	0.344
5	0.271
6	0.233
7	0.558
8	0.434
9	0.431
10	0.481

Exhibit D12: Item Factor Analysis for the *Missing Number* task

Item	Factor Analysis
1	0.124
2	0.351
3	0.197
4	-0.075
5	0.397
6	0.340
7	0.507
8	0.164
9	0.402
10	0.414

Exhibit D13: Item Factor Analysis for the *Addition Level 1 and 2* tasks

Item	Factor Analysis
1	0.248
2	0.272
3	0.049
4	0.139
5	0.151
6	0.460
7	0.374
8	0.331
9	0.269
10	0.403
11	0.496
12	0.541

Exhibit D14: Item Factor Analysis for the *Subtraction Level 1 and 2* tasks

Item	Factor Analysis
1	0.156
2	0.369
3	0.081
4	0.227
5	0.066
6	0.331
7	0.327
8	0.244
9	0.228
10	0.312
11	0.343
12	0.267

Exhibit D15: Item Factor Analysis for the *Word Problems* task

Item	Factor Analysis
1	0.142
2	0.233
3	0.302
4	0.204
5	0.227
6	0.489

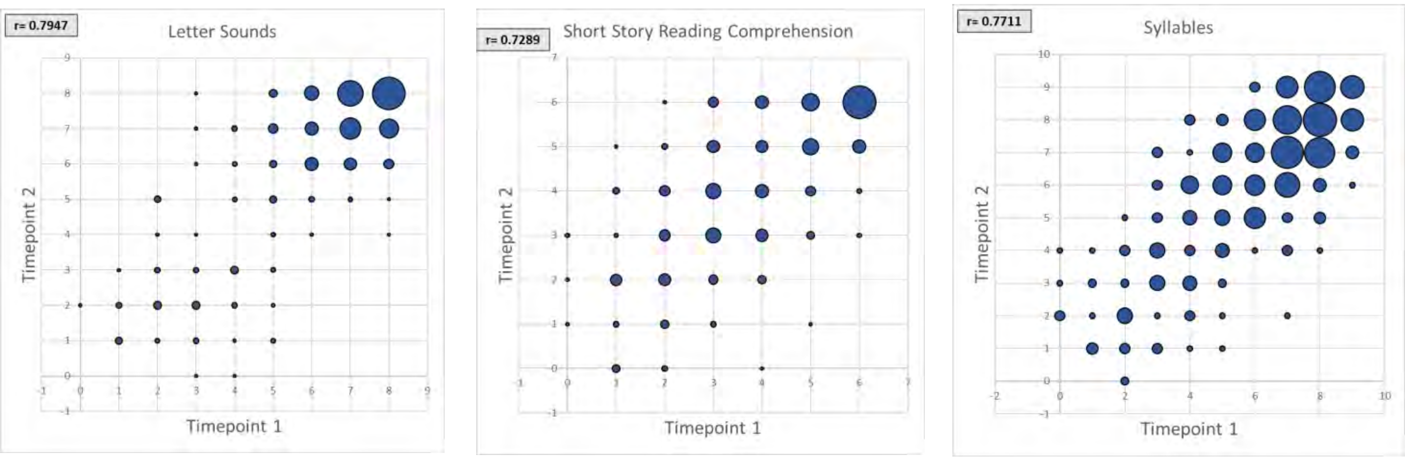
Annex E. Test-Retest Reliability: Pearson’s Correlation and Bland-Altman Plots

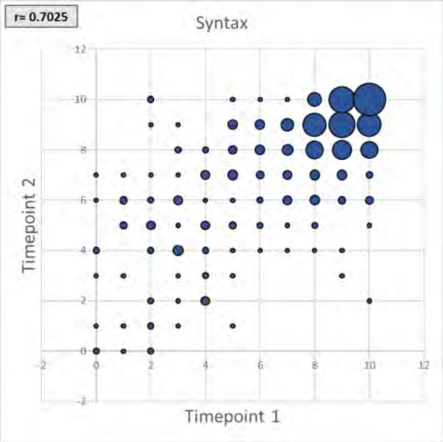
SA-EGRA

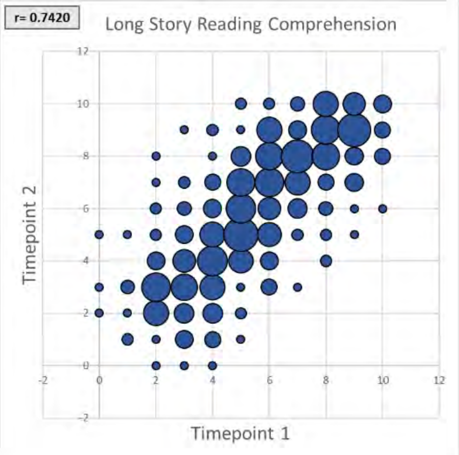
Exhibit E1: Pearson’s Correlation for the SA-EGRA Test-Retest

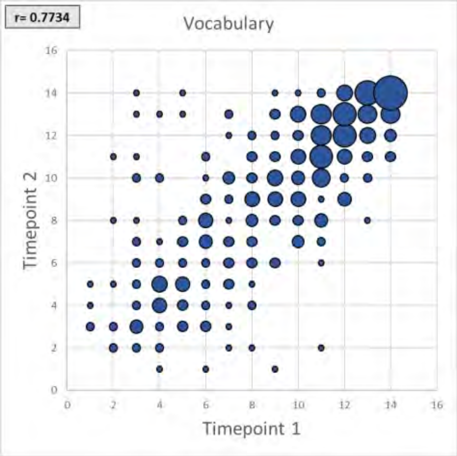
SA-EGRA Task Percent Score	Correlation
Letter Sounds	0.795
Short Reading Comprehension	0.729
Syllables	0.771
Syntax	0.703
Long Reading Comprehension	0.742
Vocabulary	0.773
Spelling	0.919

Exhibit E2: SA-EGRA Pearson’s Correlation Generalized Test-Retest Reliability









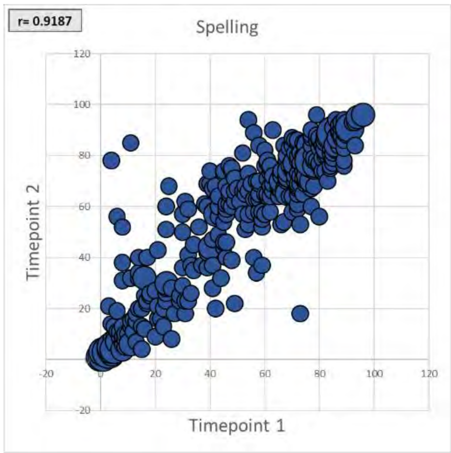
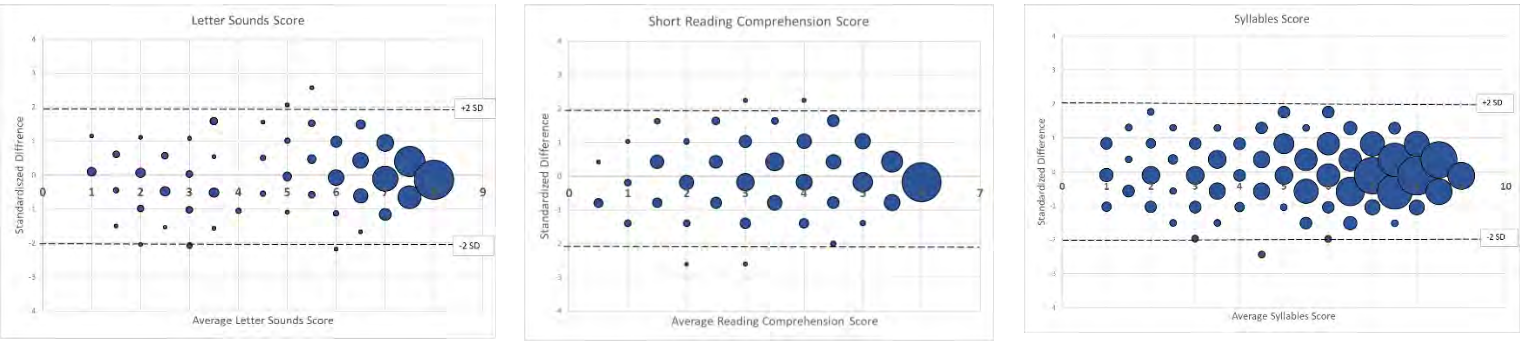
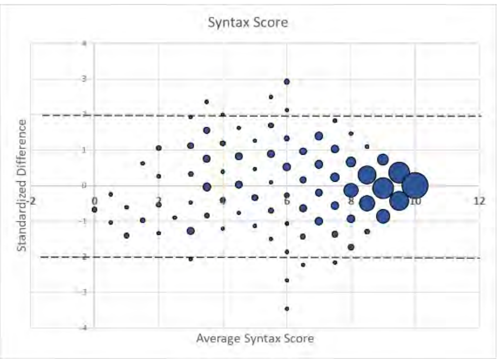
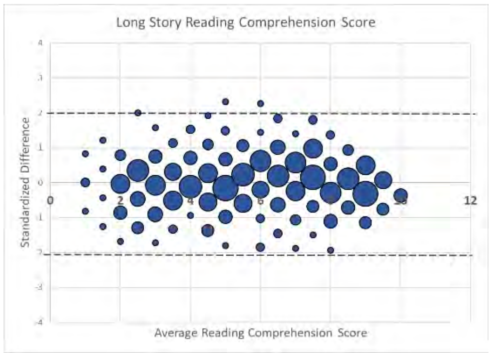
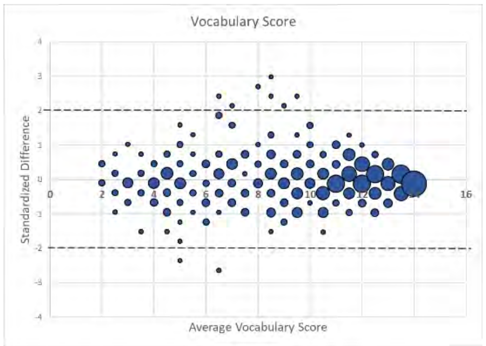


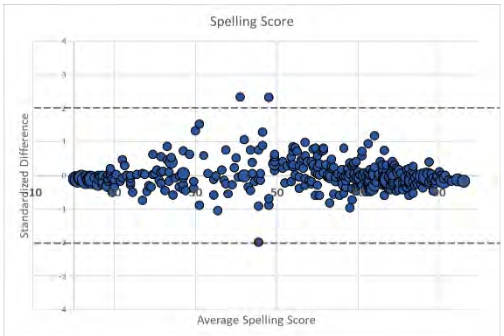
Exhibit E3: SA-EGRA Bland-Altman Plots of Test-Retest Reliability, by task











SA-EGMA

Exhibit E4: Pearson’s Correlation for the SA-EGMA Test-Retest

SA-EGRA Task Percent Score	Correlation
Number Identification	0.676
Number Discrimination	0.724
Missing Number	0.824
Addition Level 1	0.697
Addition Level 2	0.612
Subtraction Level 1	0.60
Subtraction Level 2	0.474
Word Problems	0.676

Exhibit E5: SA-EGMA Pearson’s Correlation Generalized Test-Retest Reliability

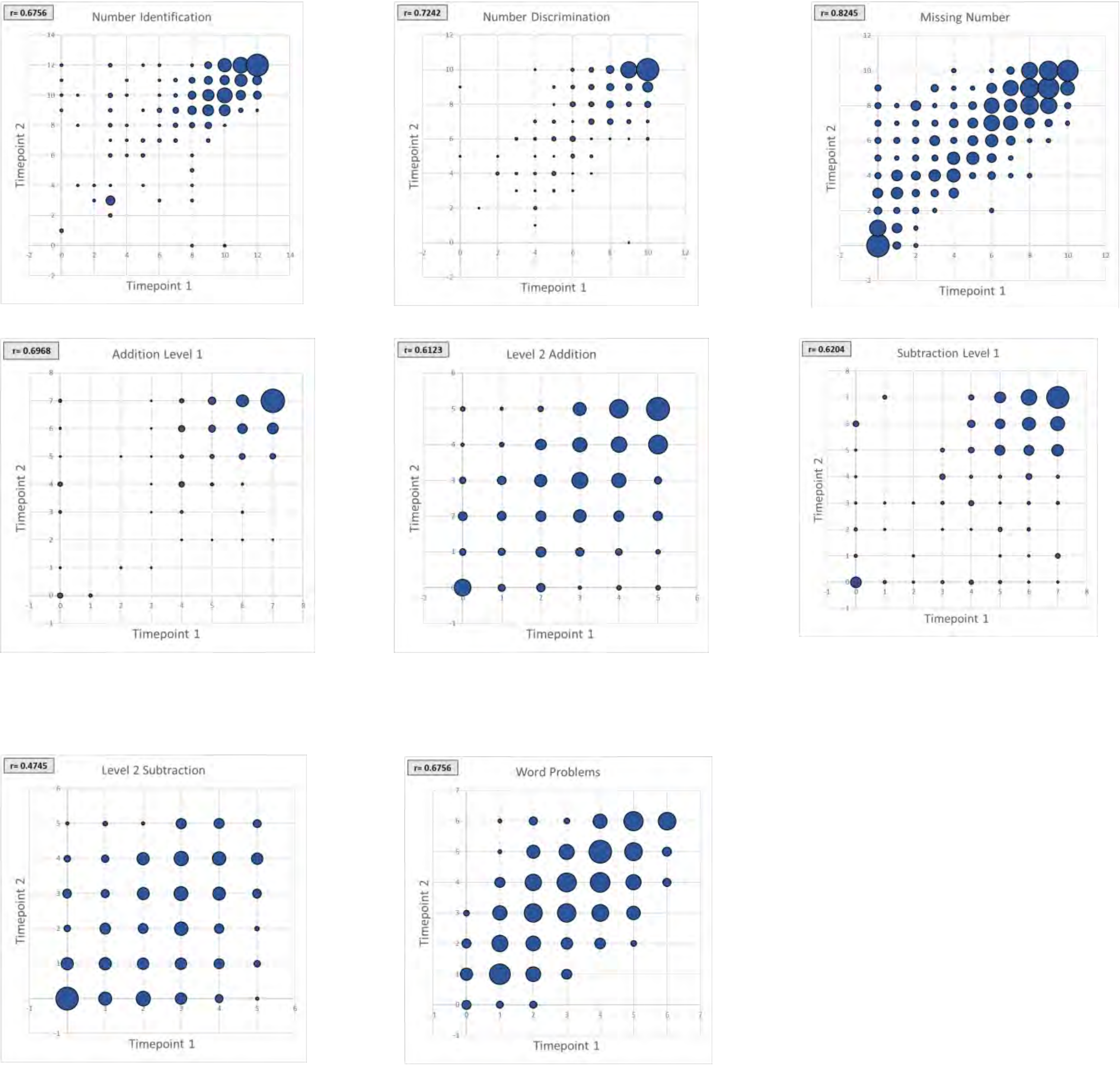
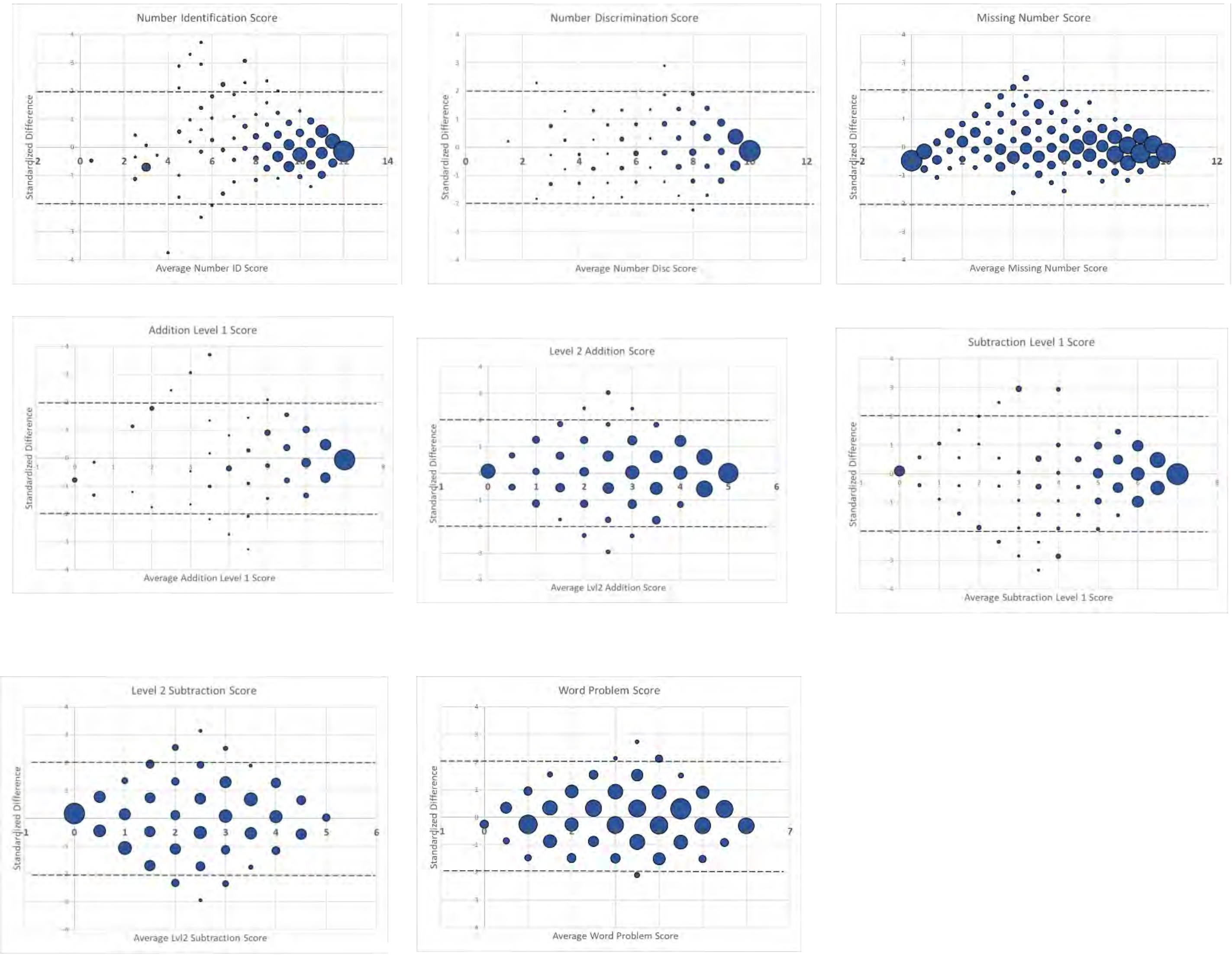


Exhibit E6: SA-EGMA Bland-Altman Plots of Test-Retest Reliability, by task


Number Identification Score

Number Discrimination Score

Missing Number Score

Addition Level 1 Score

Level 2 Addition Score

Subtraction Level 1 Score

Level 2 Subtraction Score

Word Problem Score

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

SA-EGRA

Exhibit F1: SA-EGRA Spelling Score Percent score vs. Traditional EGRA Oral Reading Fluency

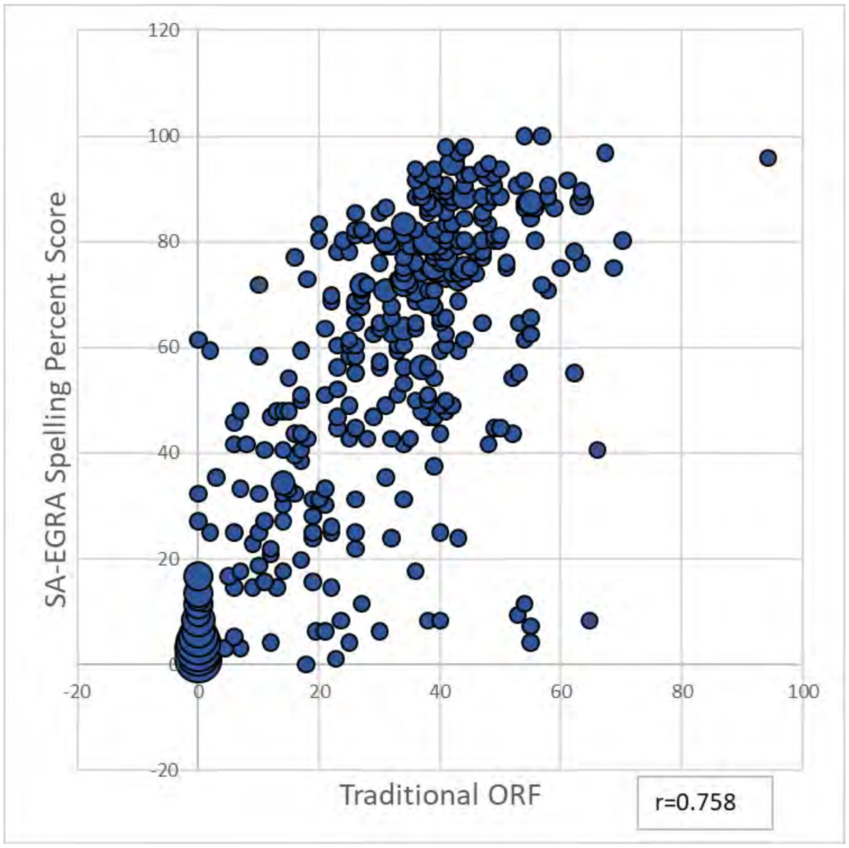
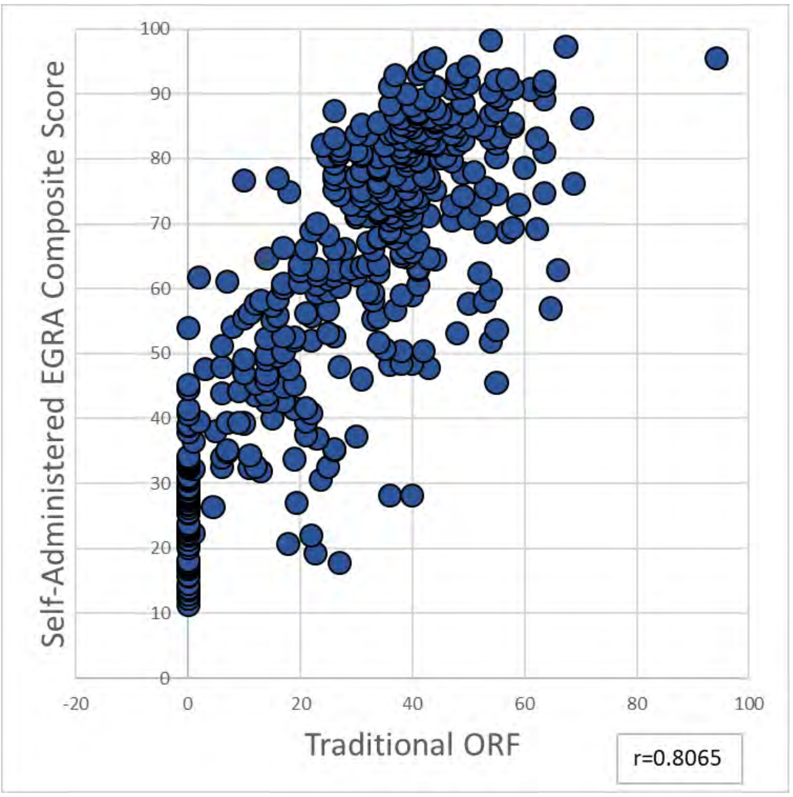


Exhibit F2: SA-EGRA Composite Score Percent score vs. Traditional EGRA Oral Reading Fluency



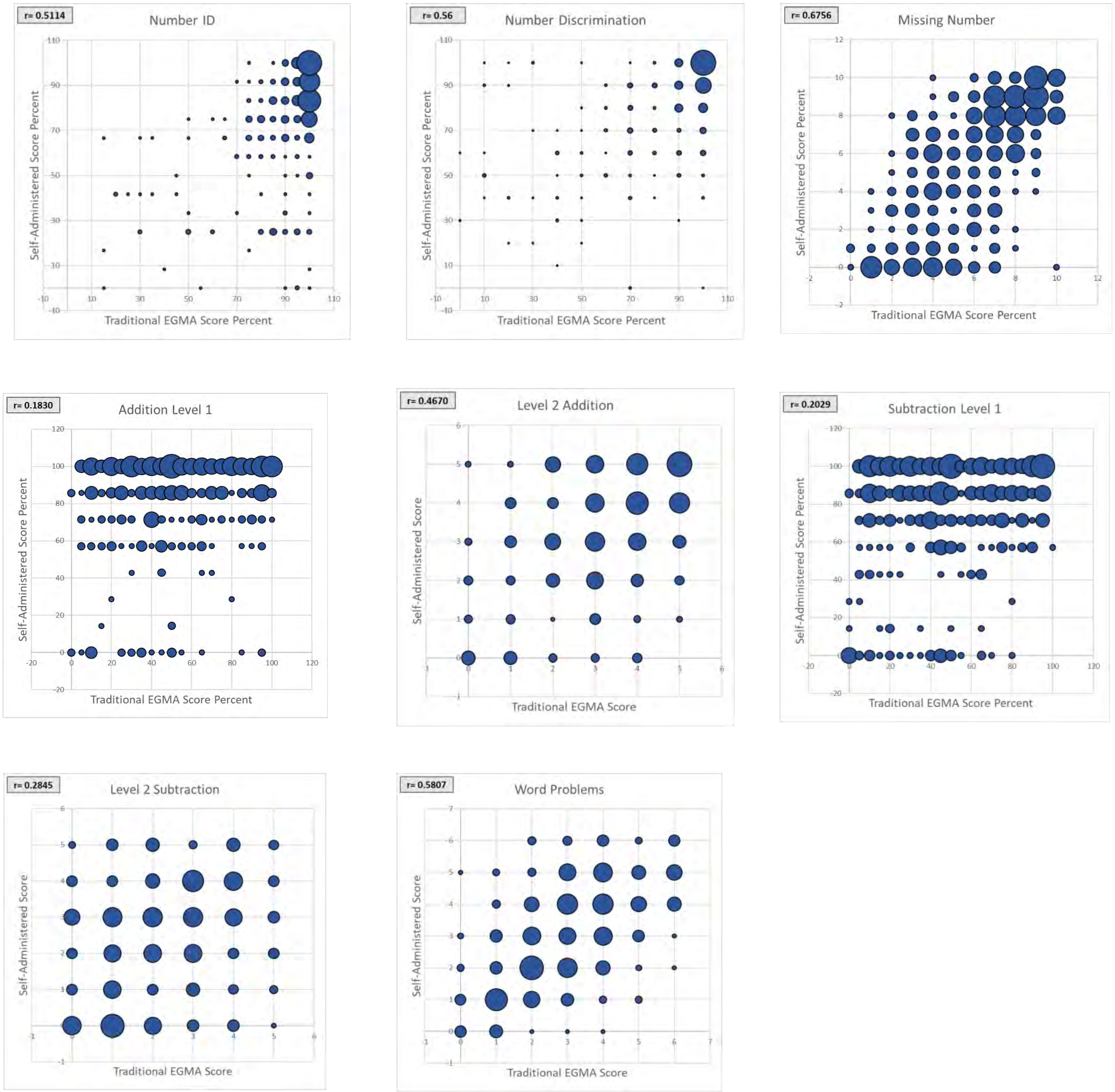
Composite score calculation: SA_EGRA_composite = 0.4*spelling_total_score_pcnt + 0.15*short_read_comp_score_pcnt + 0.1*long_read_comp_score_pcnt + 0.05*letter_sounds_score_pcnt + 0.1*vocab_score_pcnt + 0.1*syntax_score_pcnt + 0.1*syllables_score_pcnt

SA-EGMA

Exhibit F3: Pearson's Correlation for Generalized Concurrent Validity of the SA-EGMA and Traditional EGMA

SA-EGMA Task Percent Score	Correlation
Number Identification	0.511
Number Discrimination	0.560
Missing Number	0.676
Addition Level 1	0.183
Addition Level 2	0.467
Subtraction Level 1	0.203
Subtraction Level 2	0.284
Word Problems	0.581

Exhibit F4: Pearson's Correlation Generalized Concurrent Validity of SA-EGMA vs. Paper-Based


r= 0.5114
Number ID
Self-Administered Score Percent
Traditional EGMA Score Percent
r= 0.56
Number Discrimination
Self-Administered Score Percent
Traditional EGMA Score Percent
r= 0.6756
Missing Number
Self-Administered Score Percent
Traditional EGMA Score Percent
r= 0.1830
Addition Level 1
Self-Administered Score Percent
Traditional EGMA Score Percent
r= 0.4670
Level 2 Addition
Self-Administered Score
Traditional EGMA Score
r= 0.2029
Subtraction Level 1
Self-Administered Score Percent
Traditional EGMA Score Percent
r= 0.2845
Level 2 Subtraction
Self-Administered Score
Traditional EGMA Score
r= 0.5807
Word Problems
Self-Administered Score
Traditional EGMA Score