Deliverable #6:

Additional Analyses for SA-EGRA and SA-EGMA

Submitted

January 26, 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

Simon King and Karon Harden

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

Executive Summary

This report summarizes the findings of additional analyses conducted to delve deeper and develop more insight into the piloting of the Self-Administered Early Grade Reading Assessment (SA-EGRA) and the Self-Administered Early Grade Mathematics Assessment (SA-EGMA). These tools were developed and tested by RTI International with the support and direction of Imagine Worldwide.

Children complete these assessments independently on tablet-based software while in a classroom with their peers. An adult supervises the process.

This report presents additional analyses conducted on two data collections, conducted in Ghana in 2022, and designed to assess the performance of the tools. The field test assessed 421 grade one students and 429 grade three students in August 2022. The pilot study assessed 279 grade two students on the SA-EGRA and SA-EGMA, in addition to a traditional EGRA/EGMA, in September and October 2022. This report is not attempting to disseminate comprehensive psychometric findings but rather provide additional analysis and context that will inform further development and use of the SA-EGRA and SA-EGMA.

Key findings from these additional analyses include:

- Further refinement of the student self-administered fluency reading measure is necessary. While the instrument produces a range of reasonable fluency scores, the overall measure has a very low association with other student literacy subtasks, indicating the subtask is not yet measuring the construct it was designed to measure. Given the observed data patterns when comparing the subtask against the paper-based fluency measure suggest that the subtask is an accurate measure for a small percentage of students; the issue to focus on remains the self-administered design and protocol.

- The spelling subtask remains a strong proxy indicator of generalized student performance on a traditional oral reading fluency subtask, even when compared against a composite literacy score developed using Structural Equation Modelling (SEM).

- The overhaul of the syntax subtask looks to have successfully mitigated the issue of *yes* bias observed during the field test.

- The SA-EGMA number identification subtask performance decreased between the field test and the pilot test. The decrease in tool internal consistency performance at the pilot stage is likely attributed to a change in the student sample's increased numeracy skill level and the subtask unable to measure the variability when many more students successfully answered the items correctly. Note that the tool performance in the pilot study was still satisfactory.

- The above issue identified for the SA-EGMA number identification subtask is likely the issue for the decreased performance of many numeracy subtasks at the pilot stage compared with the field test. This indicates that the current SA-EGMA subtasks likely perform more optimally at lower proficiency levels.

SA-EGRA and SA-EGMA Additional Analyses

1. Introduction and Background

A tool's reliability is its ability to measure the desired construct consistently. The purpose of this activity was to conduct additional analyses to evaluate the SA-EGRA / SA-EGMA in this regard. This report on the additional analyses is intended as a companion report to the final report on results of the pilot test[1].

2. Purpose of Additional Analyses

Developing assessment tools with high validity and reliability often takes multiple refinement cycles. A data collection activity generally informs each cycle. As such, these additional analyses inform further refinement of the SA-EGRA and SA-EGMA tools.

3. Data Used for Analysis

Data for this additional analysis were collected at two-time points for different purposes, during a field test and a pilot test. These are detailed below. Children completed the assessments independently on tablet-based app, based on RTI's open-source Tangerine software platform, while in a classroom with their peers. An adult supervised the process.

3.1. Initial Field Test (August 1-5, 2022)

The field test was conducted August 1-5, 2002. Data were collected at 20 schools in the Adenta and Weija-Gbawe Municipalities, Ghana. A total of 441 grade one and 429 grade three students participated in the study. The purpose of the field test was to:

- Assess the app's rendering such that issues evident from the test administration or data analysis could be addressed.
- Assess and address any observed protocol and test administration issues.
- Assess tasks and task items for internal consistency (that the items measure the same constructs). Adapt, change, or remove tasks or items that do not meet expectations.
- Assess the duration of the assessment and remove items that are redundant in describing student literacy and numeracy skills.

3.2. Pilot Test (August 1-5, 2022)

The pilot test was conducted in Ghana from September 28, 2022, through October 11, 2022. A total of 279 grade two students participated. The purpose of the pilot was to:

- Retest the tools for internal consistency after changes were made post field-test.
- Conduct concurrent-validity component (with the same student completing both, a traditional EGRA or EGMA and its self-administered counterpart).
- Assess the test-retest reliability of the tools. Each student completed the SA-EGRA or SA-EGMA a second time 7 days after being first assessed.

[1] The final report on the pilot test can be found here: https://shared.rti.org/content/report-self-administered-egraegma-pilot-ghana-english

4. Additional Analyses Justification

There are six tasks designated for further analyses detailed below:

4.1. Exploration of the oral reading rate data

An initial oral reading fluency subtask was developed and tested during the field test. The findings at this stage suggested further changes should be made to the subtask. This exploration assesses the concurrent validity of the subtask using data from the pilot test.

4.2. Deeper analyses of the task-level durations

The SA-EGRA and SA-EGMA were both initially psychometrically assessed during the field test. Some subtasks had items removed or changed. The pilot test re-assessed these subtasks. Based on the duration of the subtasks and the performance of the items, this analysis will assess if the length of each subtask is appropriate.

4.3. Exploration of the new syntax SA-EGRA task

The initial syntax subtask struggled with *yes* bias; when presented with a yes or no choice, many students responded *yes*. While the exact reason behind students choosing *yes* is not fully known, the development team acknowledged the need to re-develop the syntax subtask and mitigate the *yes* bias.

4.4. Development of an SA-EGRA composite score using Structural Equation Modeling (SEM)

The initial report presented strong concurrent validity between students' SA-EGRA spelling subtasks score and the traditional paper-based oral reading fluency subtask. This means that there is potential for the spelling score to be used as a proxy against reading proficiency benchmarks (with caveats). This additional analysis will construct an SA-EGRA composite score, optimized using a method called Structural Equation Modeling (SEM), and see if it adds more value as a better-fit proxy model than the spelling score.

4.5. Further analysis of the number identification subtask to better understand the item-level factor loadings

The number identification subtask overall performed adequately. However, some of the items were below optimal performance levels. This analysis will attempt to isolate this issue and explain why this performance issue exists.

4.6. Further scrutiny of selected lower-loading items

There are instances of some items and subtasks underperforming. This analysis will identify a sample of lower-performing items and, if possible, quantify the reason behind this lower performance. Note that many underperforming items will be identified through the prior five analyses.

5. Additional Analyses Findings

5.1. Exploration of the oral reading rate data

The oral reading rate subtask was developed for the pilot test. The students were presented with a

61-word text and asked to read it aloud quietly to themselves and to pay attention because they would answer some questions about it when they were finished. The original version developed for the field test involved the student seeing only a portion of the text at a time and tapping on an arrow to work their way through the text; however, when this format resulted in very noisy data, the format was revised for the pilot to show the whole text at once on a static page. The reading rate was calculated as the number of words in the text divided by the amount of time lapsed between the student's initial tap to start the passage and the tap signaling that they were finished at the end.

The first step in assessing the subtask is to look at the general properties. Some students stopped the timer on reading the passage immediately without reading the passage. This resulted in unrealistic fluency rates. To account for this, we eliminated reading rates of over 150 correct words per minute (cwpm). Our pilot study removed 26.6% of student scores from further analysis. Of the remaining scores, the overall mean average was 67.8 correct words per minute (cwpm), with averages of 62.4 cwpm and 71.4 cwpm for grade two and grade three, respectively. These are reasonable given the performance of the student participants in other subtasks.

The critical step of this process is to compare the oral reading rate against the other SA-EGRA subtask scores. While we would not expect a strong positive association between (for example) reading rate and silent reading comprehension, we should expect some level of positive association. By initially assessing this association using factor analysis, we see that the oral fluency subtask lacks an association with the other subtasks (Exhibit 1) with a very low factor loading of 0.077; a minimum threshold for an acceptable factor loading is 0.3[2].

Exhibit 1: Factor Analysis Loadings for SA-EGRA Subtask Scores (reading rate<150)

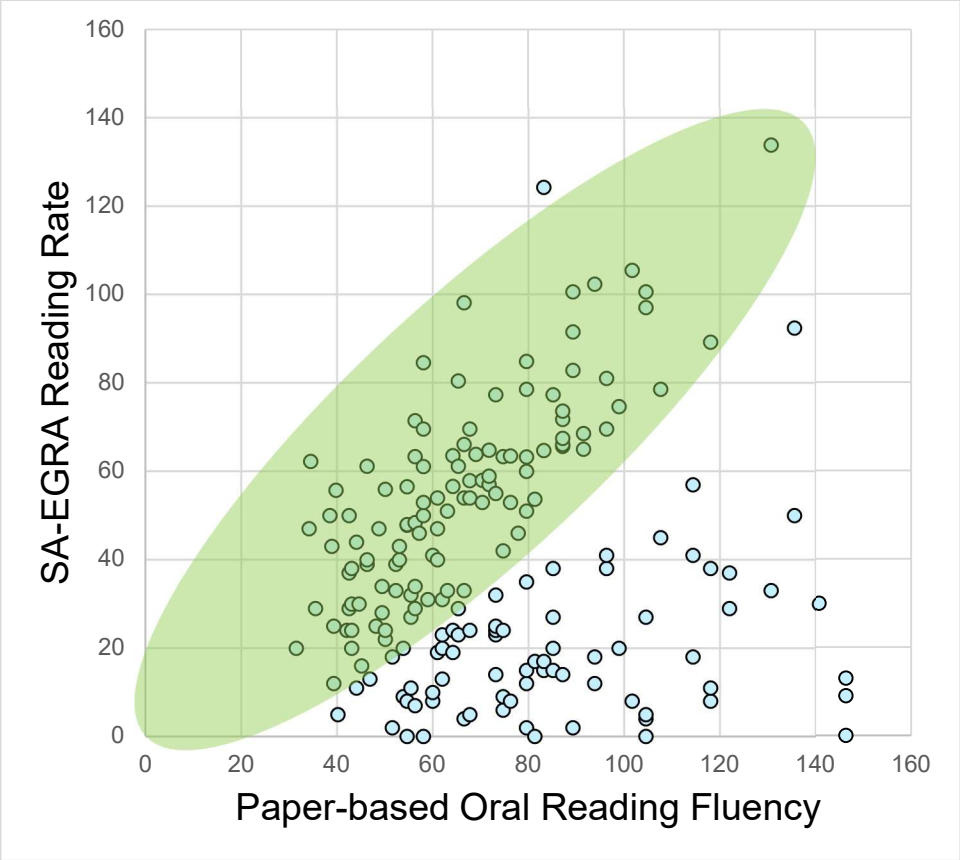
Subtask	Factor 1	Factor 2
Reading Rate	0.077	0.298
Letter Sounds	0.563	-0.188
Silent Reading Comprehension	0.621	0.103
Vocabulary	0.770	0.109
Spelling	0.820	-0.025
Syntax	0.670	0.066

Each factor represents a latent construct. The first factor could be assumed to be student literacy skills. The reading rate has a higher factor loading (0.298) for factor 2, but it is unclear what this factor represents.

The goal of developing the SA-EGRA reading rate subtask is that it performs the task of the paper-based EGRA oral reading fluency subtask, producing a student score near to how the students might score for the oral reading fluency subtask. During the pilot, students were assessed using the SA-EGRA and the paper-based EGRA. As such, we can compare their SA-EGRA reading rate to the paper-based oral reading fluency (Exhibit 2). By excluding the SA-EGRA reading rate score of over 150, we get a Pearson's correlation of r=0.149 (p=0.0426). If we include the reading rates of over 150, the correlation decreases to r=-0.221.

[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.

Exhibit 2: Scatterplot of SA-EGRA Reading Rate Versus Paper-based Oral Reading Fluency (Reading Rate<150).



By inspection of the scatterplot, there does seem to be a slight bottom-left to top-right pattern (illustrated by the ellipsoid in Exhibit 2), suggesting that some students are attempting to read the passage in the SA-EGRA version. However, given that the scatterplot already filters students scoring over 150 on the reading rate task, the number of students attempting to read the SA-EGRA passage diligently is relatively low. Therefore, the hindrance to this subtask working effectively is the self-administered tool design. On later versions of the SA-EGRA, it might be interesting to experiment with different protocols of this subtask, seeing if there was a way to nudge diligent student application to the subtask.

5.2. Deeper analyses of the task-level durations

The question of efficiency due to changes made to the subtasks between the field test and a pilot study is challenging to answer definitively. Exhibit 3 shows the average duration of the SA-EGRA and SA-EGMA tools at the field test and pilot study (two time points). Accounting for the margin of error, there is no statistically significant difference in the duration of the administration at all three timepoints (including the second data collection during the pilot study).

Exhibit 3: Timing of Assessments Between the Field Test and Pilot Study

Assessment	Mean (mins.)	Margin of Error
SA-EGRA (field test)	39.6	±3.09
SA-EGRA (Pilot Study time one)	41.5	±0.95
SA-EGRA (Pilot Study time two)	37.8	±3.08
SA-EGMA (field test)	44.4	±2.48
SA-EGMA (Pilot Study time one)	49.7	±2.36
SA-EGMA (Pilot Study time two)	41.7	±2.86

Additionally, as already explained in the pilot study report, the internal consistency of the tools decreased between the field test and pilot study even when there were no changes to subtasks. As outlined in section 5.6, the explanation of the lower internal consistency is likely contributable to assessing the tools with a student sample with higher literacy and skill levels. Therefore, it is not possible to attribute change in performance to the duration time of the overall assessment or subtasks.

To be able to study this question comprehensively, we would need to design a test-retest study where the same students attempt different tools. This type of design is necessary to control for confounding issues by making them constant (e.g., using the same students at each timepoint) and measuring just what we vary (e.g., two versions of the tool).

5.3. Exploration of the new syntax SA-EGRA task

The syntax subtask developed for the field test had students read a short phrase and comment if the statement was true. The challenge with this subtask is that many students were prone to respond that the statement was true, indicating that many were moving through the items by agreeing every time rather than deducing the correct response. The consequence was that most students correctly identified the true phrases, while also most students incorrectly identified the incorrect phrases as true. Therefore, at the field test, there was more internal consistency with students responding that the phrase was true than correctly identifying correct and incorrect phrases (Exhibit 4). The column labeled "yes" response scores shows the consistency with which students responded yes. When scored correctly, there is no internal consistency between the incorrect and correct phrases.

Exhibit 4: Field Test - Internal Consistency of Syntax Subtask

	Factor Analysis Loadings	
Item Label	**"yes" response scored as correct (even if incorrect)**	**Scored correctly**
The ball kicks the football players.	0.442	-0.473
A house has a door.	0.334	**0.318**
A goat is smaller than a chicken.	0.425	-0.412

If it grows, the flowers will rain.	0.435	-0.463
Football players kick the ball.	0.319	**0.323**
You can have a light when you see more.	0.444	-0.463
You are listening to English yesterday.	0.512	-0.522
If it rains, the flowers will grow.	0.428	**0.431**
Students pass their exams to study.	0.343	-0.390
A big is Ghana country.	0.620	-0.583
A chicken is smaller than a goat.	0.268	**0.299**
Students study to pass their exams.	0.368	**0.401**
You are listening to English right now.	0.349	**0.389**
We should cook our hands before washing.	0.427	-0.434
Ghana is a big country.	0.281	**0.296**
Morning goes to school in the children.	0.458	-0.478
You can see more when you have a light.	0.435	**0.440**
We should wash our hands before cooking.	0.422	**0.394**
A door has a house.	0.488	-0.455
Children go to school in the morning.	0.427	**0.404**

To mitigate this challenge, for the pilot study, the subtask was changed such that rather than individual phrases being presented to the students one at a time, they were presented in complementary pairs, and the student had to select the correct phrase (Exhibit 5).

Exhibit 5: Item Factor Analysis for Syntax Subtask (correct phrase bolded)

Item Number	Option A	Option B	Factor Loading
1	The ball kicks the football players.	Football players kick the ball.	0.207
2	A house has a door.	A door has a house.	0.324
3	A goat is smaller than a chicken.	A chicken is smaller than a goat.	0.196
4	If it grows, the flowers will rain.	If it rains, the flowers will grow.	0.139
5	I will write in my notebook when I find my pencil.	I will find my pencil when I write in my notebook.	0.109
6	You are listening to English yesterday.	You are listening to English right now.	0.221
7	People cut down trees to get firewood.	People get firewood to cut down trees.	0.325
8	Ghana is a big country.	A big is Ghana country.	0.486
9	We should cook our hands before washing.	We should wash our hands before cooking.	0.372
10	Morning goes to school in the children.	Children go to school in the morning.	0.311

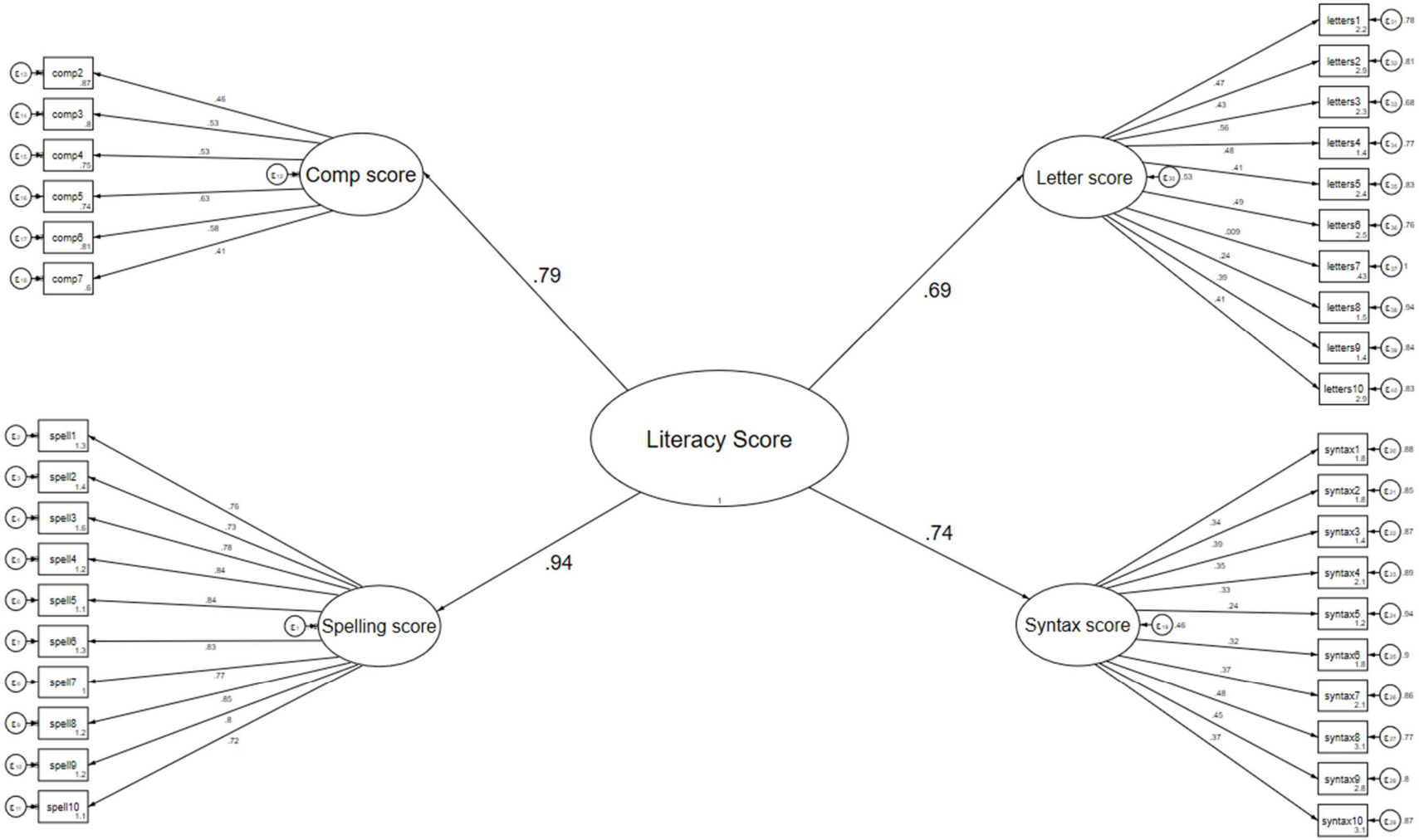
While the internal consistency of this new subtask design is not perfect (i.e., some factor loadings are less than 0.3), *yes* bias has been successfully addressed. To improve this subtask for future administrations, further refinements to the subtask might include aligning the phrases better with the expected proficiency of the students.

5.4. Development of an SA-EGRA composite score using Structural Equation Modeling (SEM)

The premise behind developing an SEM composite literacy score is that it uses a nested factor analysis approach to explore relationships and develop models. Usually, creating a composite score through factor analysis would be to conduct analyses of the subtask overall percentage scores and create a composite score using the factor weights. Generally, the subtask score that best explains student scores' variability would receive the greatest loading (or weight) to construct the linear composite score. SEM improves this process and uses a factor loading type approach to construct each subtask score and then looks to combine them into an overall literacy composite score. Additionally, composing a score using traditional Factor Analysis involves adjusting which items or scores to keep that will compose the score; SEM uses a model fit approach. Poorly fitting items will result in a lack of fit. Adapting the model or dropping the items is the best way to address this lack of fit.

The model build was attempted using all five SA-EGRA subtasks to generate a literacy composite score. The final model is shown in Exhibit 6. The standardized loadings (or weights) are shown on the connecting lines. For example, the spelling score loading for the composite literacy score is 0.94, the highest load or influence on the composite score. Comparatively, the letter sounds score has a lower loading of 0.79. The item contributions are shown with the connecting lines between the subtask scores and the subtask items. So, for example, the first spelling item has a loading contribution of 0.76 toward the composite spelling score.

Exhibit 6: SEM Model Used to Generate Literacy Score



There were several iterative steps taken to produce a model fit. Firstly, the final model does not include silent reading comprehension items 1 and 8. As explained more in section 5.6, these two items have poor factor loadings and prevent a model fit for the silent reading comprehension score. Secondly, the vocabulary subtask is missing from the model fit. It is unclear exactly why the vocabulary subtasks caused a lack of model fit. Attempts were made to include the vocabulary subtask by iteratively dropping different vocabulary items, but no combination of adding and dropping changed the lack of fit outcome.

The model fit works similarly to regression models. There needs to be an overall statistically significant model fit. This was measured through a chi-squared test which returned a score of 1028.5 (degree of freedom = 590) and a p-value of less than 0.001. However, like regression, further analysis is conducted to determine the quality of the fit. The Root Mean Square Error of Approximation (RMSEA) returned a value of 0.036. Using 0.01, 0.05, and 0.08 to represent an excellent, good, and mediocre fit[3], respectively, the literacy score model fit of 0.036 is more than acceptable.

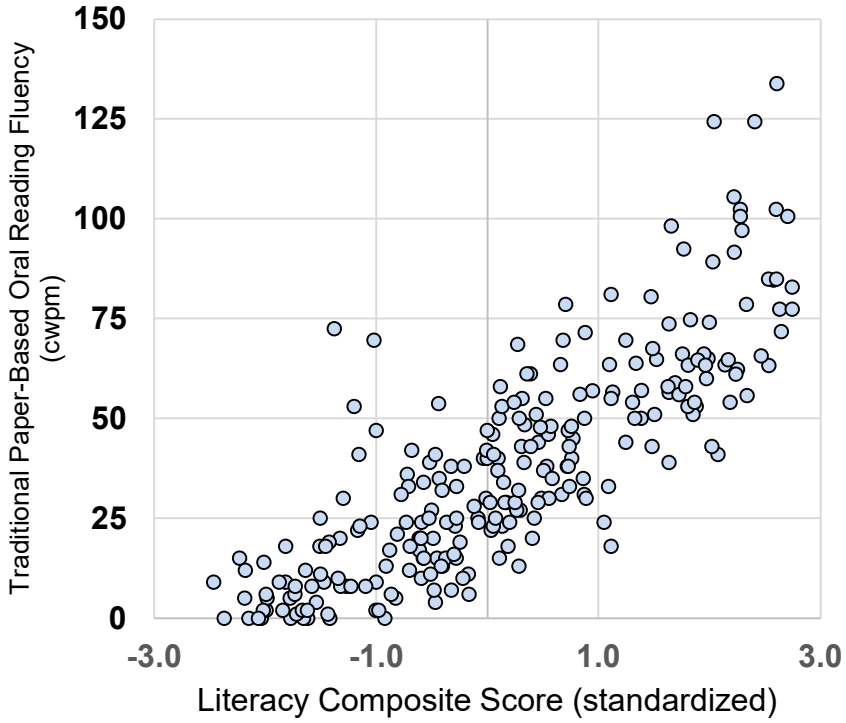
The purpose of producing this composite score was to assess options for proxy measures for oral reading fluency. As discussed in the pilot study report, the emphasis is not to use the measure as an individual student measure of reading fluency; instead, it may potentially be used for a generalized study where it was desirable to measure the percentage of students achieving a proficiency benchmark of reading fluency. Fitting the literacy composite score against paper-based oral reading fluency produces a Pearson's correlation of r=0.826 (Exhibit 7). This is a strong positive correlation. We compared this to the results from the pilot study report which fit the SA-EGRA spelling score against the paper-based reading fluency. We get a correlation of r=0.828, which is nearly identical to the literacy correlation. This similarity occurs because spelling is the largest factor loading for the literacy composite score and the spelling subtask is a particularly well-performing subtask. The spelling subtask items were each individually scored for a partial or full credit for an overall subtask potential total score of 63. This subtask could be considered the only continuous variable subtask score and is particularly good at describing more levels of student literacy skill variation than the other discrete variable subtask scores. The spelling score parallels the nature of the paper-based fluency measures, which as continuous variables, also had similar positive properties.

In conclusion, using a literacy composite score as a proxy for oral reading fluency is an option when we desire an approach that is harder to challenge from literacy and statistical perspectives. Using a literacy composite measure as a proxy is far easier to defend than a single spelling score. However, the reality is that there is little difference in performance.

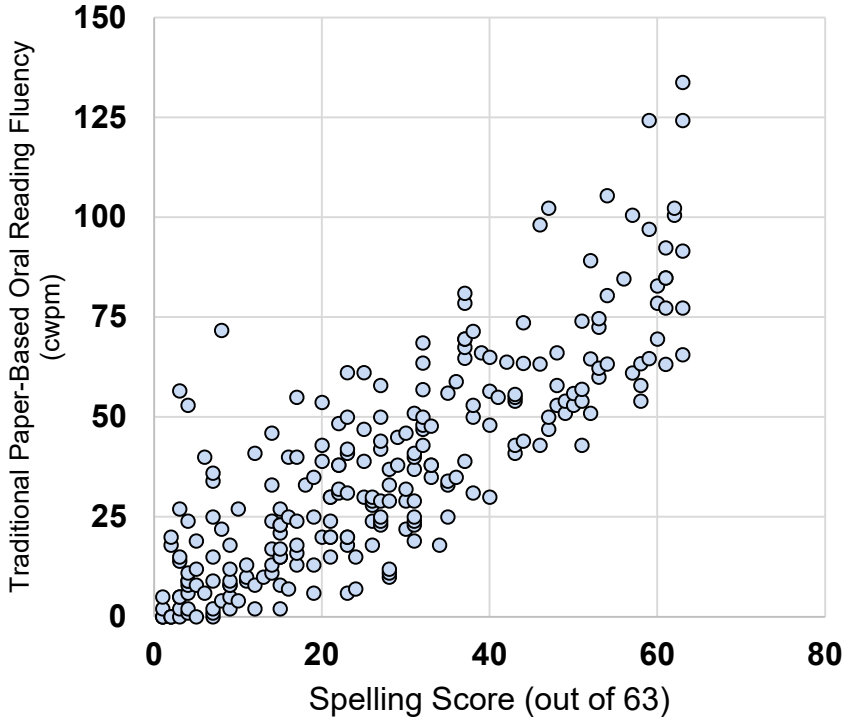
[3] MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). *Power analysis and determination of sample size for covariance structure modelling*. Psychological Methods, 1, 130-149

Exhibit 7: Scatterplots of EG-EGRA scores versus Traditional Oral Reading Fluency

Literacy Composite Score versus Paper-based Oral Reading Fluency (r=0.826)



Spelling Score versus Paper-based Oral Reading Fluency (r=0.828)



5.5. Further analysis of the Number Identification subtask to better understand the item-level factor loadings

The number identification subtask was assessed for performance during the pilot study. The internal consistency was assessed using exploratory factor analysis (Exhibit 3). The factor analysis item-level loadings are shown with the item stimulus seen by the student, and the percentage correct, incorrect, and no response (Exhibit 8). The acceptable standard for a factor analysis loading is 0.3.

Exhibit 8: Pilot Study Item Factor Analysis and Summary Findings for the Number Identification Subtask

Item	Item Stimulus	Incorrect	Correct	No Response	Factor Analysis Loadings
1	**2**	25.6	74.4	-	0.052
2	**9**	14.7	85.3	-	0.104
3	**0**	13.6	86.4	-	0.157
4	**12**	11.2	88.8	-	0.252
5	**45**	19	74.8	6.2	0.109
6	**39**	19.8	73.6	6.6	0.216
7	**80**	24.8	67.4	7.8	0.403
8	**74**	8.5	82.6	8.9	0.215
9	**66**	3.5	85.3	11.2	0.118
10	**108**	20.9	67.4	11.6	0.234
11	**587**	43	44.6	12.4	0.814
12	**989**	41.9	45.7	12.4	0.796

Unlike the pilot study findings, the item factor analysis loadings are comfortably greater than 0.4. Nine out of 12 number identification subtask items have a factor loading of less than 0.3, which is less than desirable. Seeking explanation, we observe that the two high factor loadings are when students are asked to identify three-digit numbers (i.e., 587 and 989; see Exhibit 8). The factor loadings for item 11 and item 12 are 0.814 and 0.796, respectively. We further investigated the subtasks items by reviewing the item factor loads the subtask score in the field test (Exhibit 9).

Exhibit 9: Field Test Item Factor Analysis and Summary Findings for the Number Identification Subtask

Item	Item Stimulus	Incorrect	Correct	No Response	Factor Analysis Loadings
1	**2**	58.4	41.6		0.4532
2	**9**	48.4	51.6		0.5479
3	**0**	39.5	60	0.5	0.5499
4	**12**	48.8	50.7	0.5	0.6593
5	**45**	48.8	50.5	0.7	0.6738
6	**39**	54	44.9	1.2	0.675

7	**80**	58.4	40.4	1.2	0.6464
8	**74**	39.7	59.1	1.2	0.7385
9	**66**	32.5	66.1	1.4	0.718
10	**108**	62.4	35.7	1.9	0.504
11	**587**	80.4	17.8	1.9	0.5191
12	**989**	78.3	19.9	1.9	0.4855

What also differentiates the outcomes shown in Exhibit 8 and Exhibit 9, are the percentage correct scores. The pilot study percentage correct ranges from 44.6 to 88.8, while the field study ranges from 19.9 to 66.1. This suggests that the difference in the performance of the items comes down to student sample characteristics. The field test had lower-performing students, and as a result, the number of identification items better measured the variability of student skill levels. However, the pilot study used higher-performing students who scored very well on the number identification subtask. Consequently, the subtask items for this relatively easy subtask struggled to differentiate student skill levels, resulting in lower factor loading scores. The main signal for this is that students were mostly able to identify one- and two-digit numbers for the pilot study, but many struggled with the three-digit numbers. Thus, these two items were better able to differentiate student skill levels.

In conclusion, while the number identification subtask performs well, it loses its effectiveness when students have mastery of identifying 1- and 2-digit numbers. If the SA-EGMA was to be used to assess higher-performing students, the items should include only 2- and 3-digit numbers to be identified.

5.6. Further scrutiny of selected lower-loading items

There are instances of some items and subtasks underperforming. This is most easily identified through internal consistency factor analysis of a suite of subtask items. Earlier analyses in this report have identified instances where this was the case.

Section 5.5 investigated the less-than-optimal performance of the number identification subtask, concluding that the issue was the higher performance of the pilot student sample compared with the field test student sample. This performance difference impacted how effectively the number identification subtask was able to describe variability in student performance .

The pilot sampled students found many of the SA-EGMA subtasks easier to accomplish than their peers who took the SA-EGMA at the field test. This issue is not an isolated occurrence. Exhibit 10 shows the internal consistency factor analysis for the missing number subtask at the field and pilot test stages to illustrate this issue repeating. While the subtask performed adequately at both stages, the subtask is a more optimal discriminator of student ability when the students have less mastery, in this case, at the field test, than for higher-performing samples.

Exhibit 10: Missing Number Internal Consistency Factor Analysis
Field Test versus Pilot Study

Item	Field Test	Pilot Study
1	0.797	0.330

2	0.787	0.276
3	0.860	0.375
4	0.817	0.004
5	0.745	0.531
6	0.59	0.139
7	0.632	0.488
8	0.786	0.414
9	0.744	0.554
10	0.331	0.179

This issue then likely snowballs into a problem for internal consistency at the subtask score level (Exhibit 11).

Exhibit 11: Factor Analysis Loadings for SA-EGMA Subtask Scores Field Test versus Pilot Study

Subtask Score Percent	Field Test	Pilot Study
Number Identification	0.542	0.355
Number Discrimination	0.583	0.438
Missing Number	0.797	0.652
Addition	0.775	0.544
Addition Level 2	0.834	0.533
Subtraction	0.837	0.462
Subtraction Level 2	0.754	0.557
Word Problems	0.616	0.534

The composite literacy score development described in Section 5.4 included dropping the silent reading comprehension item 1 and item 8. These items had factor loadings of less than 0.1. The first item asks, "*In the story, what did Esi do for the first time?*". Only 21% of the students correctly responded, "Visit the city." The incorrect response, "Get up early," was selected by 57%. This question is more challenging than items 2-7, which asked more straightforward questions such as "What color were Esi's laces?" and "What does Esi wear?". Consequently, while the first item does not harm the assessment, it does not add value in its ability to differentiate between student ability levels. However, it has merit as a diagnostic question to assess individual or group comprehension progress.

Letter sounds item 7 has been challenging at both the field test and pilot stages. The sound provided by the audio recording is /n/. For the pilot, 76% of the students responded with /m/, and only 17% responded correctly with /n/. This was observed at the field test stage, and the audio (the speaker is Ghanaian) was re-recorded for clarity for the pilot. However, students still found this item challenging. As the SA-EGRA is translated and tested in other settings and local languages, if this item is used again, it will be useful to track its performance to isolate the issue why students consistently found this item challenging.