SENSE PROJECT EVALUATION:
READING PROFICIENCY ASSESSMENT

RESULTS OF THE HAUSA EARLY GRADE READING ASSESSMENT (EGRA) IN ADAMAWA AND GOMBE

October 26, 2021

This report was produced by Katharina Hammler, PhD on behalf of AUN for review by the United States Agency for International Development Nigeria (USAID Nigeria). The opinions expressed therein do not necessarily reflect those of USAID.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. EXECUTIVE SUMMARY

This report aims to assess **changes in students' basic literacy** in an effort to evaluate the second cohort of the SENSE literacy program. Data to assess Hausa literacy levels of SENSE project participants was collected in Gombe and Adamawa State for the baseline in November of 2020, and for the end-of-project assessment in July of 2021. The assessment was done using the Early Grade Reading Assessment (EGRA) tool, consisting of seven subtasks:

- Letter identification (correct letter sounds per minute, clspm)
- Syllable identification (correct syllable sounds per minute, csspm)
- Familiar word reading (correct words per minute, cwpm)
- Non-word reading (correct non-words per minute, cnonwpm)
- Oral reading fluency (correct words per minute in a connected text, orf)
- Oral reading comprehension
- Listening comprehension

Assessment **data is available from 1,929 grade-2 students** (965 from the baseline, and 964 from the end-of-project). All students received a full year of the SENSE intervention. The research design allows only to track learning outcomes; it does not allow to establish a causal contribution of the SENSE program.

The end-of-project assessment shows that **40.9% of students in the SENSE program attain a minimum grade-level proficiency in reading at the end of grade 2** (USAID performance indicator **ES 1.1.**), up from 3.34% at the baseline. The estimated percentage of learners targeted for USG assistance with an **increase of at least one proficiency level in reading at the end of grade 2 (ES.1-48) is 63.5%.** No differences between boys and girls could be established for these indicators. These results are obtained by mapping the EGRA results onto the Global Proficiency Framework (GPF) for literacy. The available data allows to carry such mappings out for the GPF subdomains *Aural Language Comprehension*, *Decoding*, and *Reading Comprehension*. This mapping also reveals that at the end of the project, only about 5.9% of students were not meeting or only partially meeting the minimum threshold for aural comprehension (down from 36.2% at baseline). In the decoding subdomain, the respective percentage of students decreased from over 95% at the baseline to about 57.5%. In the subdomain reading comprehension, the respective statistic decreased from about 90% at the baseline to about 36% at the end-of-project assessment.

The report reveals **considerable improvements in all analyzed EGRA subtasks, both for boys and girls.** The changes between survey rounds are highly statistically significant. The data shows that on average, at the end of the project students were able to correctly identify 39.7 (baseline: 4.27) letter sounds per minute; to read 36.15 (baseline: 2.47) syllable sounds per minute; to read 23.15 (baseline: 1.48) familiar words per minute; to decipher 18.55 (baseline: 1.03) invented words per minute; and to correctly read 28.39 (baseline: 3.49) correct words per minute in a connected text. Additionally, students were on average able to answer 3.1 (baseline: 0.53) out of 5 reading comprehension questions and 4.39 (baseline: 3.0) out of 5 listening comprehension questions correctly.

A review of **potential student characteristics correlated with the size of learning gains** found that of all reviewed factors what mattered was the state of residence (living in Gombe). Additionally, student coming from more disadvantaged backgrounds and those speaking Hausa at home seem to increase their skills the most. Finally, an analysis of **complementary data** qualitatively supports the assertion that the SENSE program contributed to these improved literacy outcomes: at project end, most students thought their reading skills improved a lot (67%) or a bit (27%) over the past six month, up from only 17% and 26%, respectively, at the baseline. Students also reported increased enjoyment of reading, and were more likely to think that the reading materials available to them are helpful.

2. INTRODUCTION

2.1. PROJECT BACKGROUND

Insurgency for nearly a decade coupled with other challenges has brought the educational system in northeast Nigeria to nearly total collapse. The American University of Nigeria (AUN) with funding from United States Agency for International development (USAID) is implementing the Strengthening Education in Northeast Nigeria (SENSE) Project targeting Adamawa and Gombe States starting March 27th, 2019 through March 26th, 2022.

The project approach is to develop and strengthen the education system to enable it to deliver access to high quality education options that respond to the needs of all children without burdening it with additional cost that will result if parallel systems are created. The project works with state partners (State Ministry of Education, SUBEB, LGEA, Colleges of Education, States Universities and national and international academic institutions to analyze and understand the situation, prepare strategies for various scenarios and test and apply the best fit for purpose in each situation. The project aims to constantly measure the effectiveness of each of these approaches to identify the best approaches for scale-up and replication.

It is hoped that the project will in three years:

- Create an improved education management capacity of 100 education managers, emphasizing female leaders;
- Reach 200 primary schools;
- Improve the teaching skills of 5,000 teachers and provide skills and support for female teachers; and
- Improve educational outcomes for 200,000 primary school children, i.e., for both boys and girls.

2.2. PROJECT EVALUATION STRATEGY

In order to be able to assess progress towards and success in the project objective of improving educational outcomes for 200,000 primary school children, an assessment schedule was created. This original schedule provided for baseline and end-of-project assessments of each program cohort, as well as for concurrent testing in non-participant schools in order to determine the SENSE program’s contribution to any measured learning improvement using a case-control strategy. However, the COVID-19 pandemic has caused some changes to the initial project design and assessment schedule. Few months after the first baseline Hausa EGRA assessment had been conducted, Nigerian public primary schools closed down, so that the SENSE program could not proceed as initially planned, and the assessment schedule had to be adapted. Specifically, no follow-up assessment was conducted for the first project cohort, and the plan to assess students in non-participating schools had to be discarded. In participating schools, the testing schedule ended up being as follows:

- At the baseline, a first large-scale numeracy and literacy assessment was carried out with 57,473 students from 669 schools between July and August 2019, and an extensive report on the results is available (AUN 2019, 20).
- A subsequent baseline assessment of the Hausa literacy level of learners in the first project cohort in the two project states was conducted between January and March 2020; a report of the results is also available (AUN 2020a).
- A baseline assessment for the second project cohort was conducted in November of 2020 (results reported in AUN 2020b).

- This second project cohort was assessed again at the end of the school year, in July of 2021, and a report on the assessment results was prepared (AUN 2021).

This present report compares baseline and end-of-project assessment results of the second cohort in an effort to evaluate the learning gains made by participants in the SENSE program. Due to the lack of a comparison group, the comparison does not allow to make any *causal* claims about the project's contribution to any observed improvements. In fact, given that students were assessed at the beginning and at the end of grade 2, improvements are to be expected independently of the SENSE activity. The report draws mostly on EGRA assessment data, and complements the findings with an analysis of student background characteristics that may influence the results, as well as with data on program and learning satisfaction reported by program participants. A concluding discussion draws on qualitative evidence to assess the plausibility of a causal contribution of the SENSE approach.

2.3. THE EGRA TOOL

Students' Hausa literacy is assessed using the Early Grade Reading Assessment (EGRA) tool. EGRA was developed to measure student literacy learning in early grades in low-income countries, as no simple, effective, and low-cost measure for student learning had been available for this context (RTI International 2009). Today there are several versions of the EGRA tool (RTI International 2016); for the assessment at hand, a version of the NEI+ Hausa EGRA instrument was used (USAID 2011). Each EGRA survey consists of a set up subtasks that aim to measure different components of literacy learning. The EGRA tools used for this project contain the following subtasks (with the proficiency indicator in parenthesis):

- Letter identification (correct letter sounds per minute, clspm)
- Syllable identification (correct syllable sounds per minute, csspm)
- Familiar word reading (correct words per minute, cwpm)
- Non-word reading (correct non-words per minute, cnonwpm)
- Oral reading fluency (correct words per minute in a connected text, orf)
- Oral reading comprehension
- Listening comprehension

For each subtasks, proficiency scores are calculated separately; there is no overall score or proficiency classification. Instead, the authors of the EGRA toolkit suggest that national benchmarks be developed (RTI International 2016). No such official national benchmarks are available for Nigeria. Thus, the results of the assessment will be first analyzed without reference to external benchmarks, and subsequently mapped on the proficiency levels defined in the Global Proficiency Framework (see section 5).

3. DATA

3.1. DATA COLLECTION

3.1.1. TARGET POPULATION AND SAMPLING METHODOLOGY

The SENSE Activity used a three-stage process to select the baseline and end-of-project EGRA sample: (i) selecting representative sample size of learners from the total population of grade 2 learners in the 335 intervention schools, (ii) selecting the sample of schools, and (iii) selecting the sample of learners for the baseline and end-of-project assessment.

3.1.1.1. STAGE 1: DETERMINATION OF SAMPLE SIZE

The SENSE project aims to reach 200,000 learners over its project duration. For the purpose of this assessment, this number was divided equally between the two states Adamawa and Gombe, so that each state was allocated 100,000 learners to reach. Based on this number of learners to be reached by each state, that the sample size for the assessment was determined using 95% confidence level and a confidence level of 5, resulting in a sample size of 382 learners for each state. A further 100 learners were added in each state in other to account for potential attrition before the end-of-project assessment, as well as to ensure sufficient statistical power for an impact analysis that will take potential school or district clustering effects (similarities of students within schools or LGAs) into account. This sample size was distributed across the targeted schools in Adamawa and Gombe States respectively. Table 1 shows sample size distribution by state.

	State	
	Adamawa	**Gombe**
Leaners to be reached for the Activity duration	80,000	120,000
Sample size of learners (using 95% confidence level)	382	382
Addition to the sample size to account for attrition	100	100
Total representative sample of leaners to be tested	482	483

Table 1 Overview of sample size distribution

3.1.1.2. STAGE 2: SELECTING OF SAMPLE OF SCHOOLS

The SENSE Senior M&E Officer selected 222 schools from the 335 intervention schools to be a representative sample for both baseline and end-of-project assessment for the two states. The sampling procedure ensured that all the LGAs had an adequate number of sample schools to provide reliable and valid data about learners' reading learning outcomes. The sample size distribution across LGAs can be found in Annex I.

Students in the same schools were to be assessed both at baseline and at the end-of-project assessment. However, due to heavy flooding in some areas, some schools could not be reached for the end-of-project assessment and were thus replaced with randomly selected schools in the same area.

3.1.1.3. STAGE 3: SELECTING OF SAMPLE OF LEARNERS

The number of learners in the intervention school selected for the assessment was determined using the percentage of grade 2 learners in the respective schools, distributing the sample size proportionally across schools. The assigned sample size of learners in the sampled schools used in the first round of the assessment was maintained and used in the second round of the assessment. In each school, out of all grade 2 students, participants for the assessment were selected at random using a random number generator to get the total number of learners defined for each school.

3.1.2. RECRUITMENT AND TRAINING OF EGRA ENUMERATORS AND SUPERVISORS

3.1.2.1. RECRUITMENT

Independent enumerators were used during the two rounds of data collection. They were recruited using the below listed criteria:

a. Possession of at least National Certificate of Education (NCE)
b. Prior experience of administering EGRA survey.

c. Ability to read and speak Hausa and English fluently
d. Previous experience of field data collection
e. Experience of using a computer or hand-held electronic device (tablet, smartphone).
f. Interrater reliability score at the end of enumerator training

The supervisors used in both rounds of the assessment were Quality Assurance Officers and are staff of SUBEB in the target LGAs were the assessment took place in both the two states.

3.1.2.2. TRAINING

In each round of data collection, enumerators and supervisors received three days intensive training on EGRA and its administration using the tangerine application. Three days were used to train the enumerators and supervisors because they had been trained the previous year 2019 and so they had prior experience of administering EGRA during the first English EGRA and EGMA that was discarded as a result of SENSE Activity change in focus from system strengthening to improving reading outcomes in Hausa language. After the Activity refocus, the baseline training of enumerators took place from 7- 9th October 2020 and end-of-project training took place from 10-12th July, 2021. Fifty four (54) independents enumerators attended the training in each round of data collection.

At baseline data collection, former Deputy Chief of Party for SENSE Activity and reading expert Dr. Grace Malgwi facilitated the training of enumerators and project staff in Adamawa and Gombe states separately. At the end-of-project, the trained Activity M&E Staff facilitated the training (Based on RTI Guidance Notes for Planning and Implementing EGRA) concurrently in Adamawa and Gombe states. Thereafter, the independent consultant hired took charge of the data collection process for unbiased assessment. During baseline and end-of-project trainings of enumerators and supervisors, the contents of the training consisted of the following activities:

a. Review the EGRA principles to gain a comprehensive understanding of the EGRA instrument components;
b. Practice EGRA administration and scoring procedures;
c. Practice conducting the EGRA on tablets (using Tangerine software);
d. School visit to none EGRA participating school to field-test the instrument;
e. Roles and responsibilities of both enumerators and supervisors in the field;
f. Interrater reliability (IRR) administration and scoring evaluation.

In each rounds of the training, all the enumerators underwent Interrater reliability tests in order to ensure the reliability of scoring between enumerators. Only enumerators with score of at least 90 percent and above in IRR test were hired for the actual data collection.

3.1.3. ASSESSMENT ADMINISTRATION

The EGRA in Hausa instrument used for the two rounds of the assessment was adapted from NEI+ Nigeria that implemented USAID funded early grade reading project in Sokoto and Bauchi states. The assessment was conducted using the Tangerine software. Each enumerator was given a target number of pupils to assess in each school, with a student identification form to fill in the details of each pupil assessed as well as a tablet, pupil stimuli and an instruction protocol to aid the activity.

In each round of the assessment, the independent trained enumerators administered the assessment to the sample of learners in schools in over five day's duration in Adamawa and Gombe states using the learner sampling method taught during the training and EGRA Protocol and Pupil stimuli. The enumerators used interval-sampling method to select learners for the assessment at both baseline and end-of-project; the enumerators preferred this method because only learners that are present at the school on the day of the assessment are selected for the assessment. During the administration of the assessment, it took an enumerator approximately 30 minutes to administer the EGRA assessment per learner, with 15 minutes used for socio-demographic questions and the other 15 minutes for actual EGRA assessment. In one day, an enumerator could assess approximately six learners, depending on the distance of travel to the sampled schools.

Baseline data was collected by the SENSE project team in November 2020, while end-of-project data was collected by an independent consultant in July 2021. In addition to the EGRA Hausa assessment, some socioeconomic student background data was collected.

3.1.3.1. SUPERVISION OF ENUMERATORS

The trained LGA Quality Assurance Officers supervised the enumerators during the administration of the assessment to learners in schools to ensure the process of administering EGRA assessment is follow correctly. In addition, on daily basis, when all the learners' responses were synchronized to the central tangerine server, the second lead consultant who also supervised the administration of EGRA schools reviewed the uploaded data to:

1) check for errors and ensures the grade was correctly assessed,
2) ensure the enumerators assessed the correct number of learners from the correct, sampled schools, and
3) check for any other inconsistencies.

3.2. DATA CLEANING AND PROCESSING

3.2.1. INITIAL DATA CLEANING

The data from the Tangerine server was downloaded and cleaned by the second lead consultant Dr.Jamiu Olumore in readiness for data analysis. This second cleaning was guided by the following checklist:

1) Review incomplete assessments;
2) Remove any "test" assessments that were completed before official data collection began
3) Ensure that all assessments are linked with the appropriate school information for identification
4) Ensure child's assent was both given and recorded for each observation
5) Ensure that all timed subtask scores fall within an acceptable and realistic range of score.

The cleaned data sent was then sent to the international M&E Consultant-Dr. Katharina Hammler for data analysis and report writing.

3.2.2. FURTHER DATA PROCESSING

The independent international M&E Consultant Dr. Katharina Hammler then imported into Stata13 for further analysis. Standard EGRA indicators were calculated following the guidelines of the EGRA toolkit (RTI International 2016). In a further round of data cleaning, observations were excluded from the analysis if results across standard EGRA indicators were clearly inconsistent (e.g., a correct-words-per-minute score above 180, or a top orf score when the score for foundational skills was zero). In this process, around 1% of observations were removed.

3.3. DATA SET AND DESCRIPTIVE STATISTICS

Data is available from two rounds of data collection: The baseline assessment was carried out in November 2020 with 965 learners (482 from 146 schools in 11 LGAs in Adamawa, and 483 from 69 schools in 11 LGAs in Gombe). The end-of-project assessment was done in July 2021 with 964 learners (481 students from 125 schools in 11 LGAs in Adamawa, and 483 students from 70 schools in 11 LGAs in Gombe). The results of these assessment are described in detail in the baseline and end-of-project report, respectively, so that this evaluation report will focus on a comparison between the survey rounds. Data is analyzed taking into account the sampling design (school strata, and sampling probability weights).

Table 2 summarizes the study population and sample for both survey rounds. The first part of the table (columns 1-4) presents an overview of the sample, reporting the number of boys and girls and the average age of all students tested in each State and round. The baseline sample contains a total of 457 boys and 508 girls, with an average age of 7.8 years. The end-of-project sample contains a total of 471 boys and 493 girls, with an average age of 8.8 years (see also left pane of Figure 1).

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
		Sample				Survey population				
		Gender			Age	Gender		Age		
		male	female	total		Male	Female	Mean		
		n	n	n	mean sd	row proportion linearized se		mean linearized se	95% Conf. Int.	
Gombe	Baseline	231	252	483	7.8 1.3	0.48 0.02	0.52 0.02	7.82 0.06	7.70	7.93
	End-of-project	236	247	483	8.8 0.9	0.50 0.02	0.50 0.02	8.84 0.04	8.76	8.93
	Total	467	499	966	8.3 1.2	0.49 0.02	0.51 0.02	8.31 0.04	8.22	8.39
Adamawa	Baseline	226	256	482	7.8 1.3	0.47 0.03	0.53 0.03	7.81 0.06	7.69	7.93
	End-of-project	235	246	481	8.8 1.0	0.51 0.03	0.49 0.03	8.76 0.05	8.65	8.86
	Total	461	502	963	8.3 1.2	0.49 0.02	0.51 0.02	8.28 0.04	8.20	8.37
Total	Baseline	457	508	965	7.8 1.3	0.48 0.02	0.52 0.02	7.81 0.04	7.73	7.90
	End-of-project	471	493	964	8.8 0.9	0.50 0.02	0.50 0.02	8.80 0.03	8.73	8.87
	Total	928	1001	1929	8.3 1.2	0.49 0.01	0.51 0.01	8.29 0.03	8.24	8.35

Table 2 Description of dataset

The second part of the table presents the characteristics of the survey population, taking into account the survey design (sampling weights and school strata). Columns 5 and 6 show the percentage of boys and girls estimated to attend grade 2 in each state, for the study population. The data suggest that 49% of students are boys, and 51% are girls; the differences in the estimates between the survey round are within the statistical error range (see also Figure 1, right pane). Column 7 gives an estimate of the average age of students in the study populations' grade 2 classes in each state. The average age at the baseline survey is estimated as 7.81 years (7.81 for Adamawa, and 7.82 years for Gombe). At the end-of-project assessment, the estimated age has as expected increased by a year and is 8.80 years (8.76 for Adamawa, 8.84 for Gombe). Column 8 gives the 95% confidence intervals for the age estimates. Unless otherwise indicated, all data analysis presented in the rest of the report uses survey estimation techniques with probability weights and school strata.

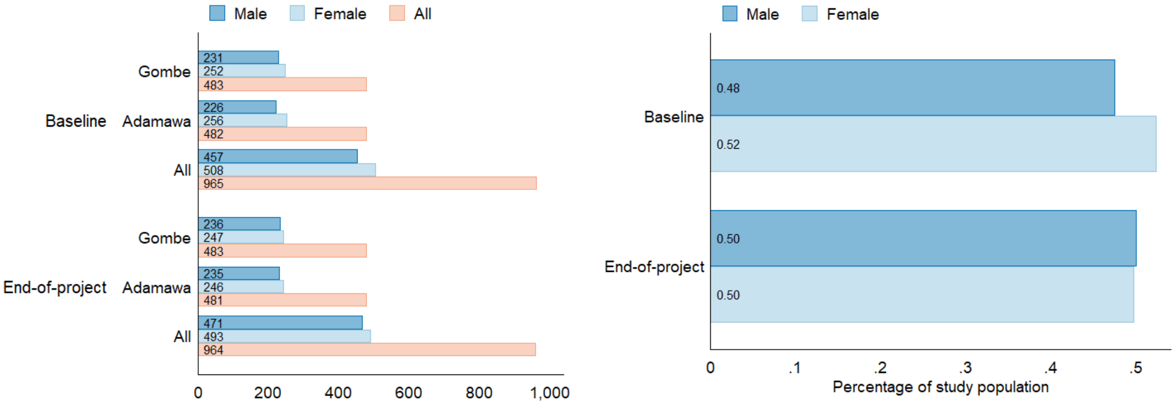


Figure 1 Left pane: Number of students tested per round, by state and gender. Right pane: Estimated percentage of boys and girls in the study population, by survey round

4. EGRA RESULTS

Overall, the two rounds of EGRA assessments reveal a strong increase in Hausa reading proficiency. As summarized in Table 3, measured reading proficiency has increased across all subtasks, with the improvement rates being both substantial and statistically highly significant. At the baseline, learners were, on average, able to correctly identity 4.27 correct letter sounds per minute; by the time of the end-of-project assessment, this score improved to 39.71 correct letter sounds per minute, representing a change of over 800%. For the subtask of syllable sound identification, the respective values are 2.47 at baseline and 36.15 at the end of the project (a change of over 1,300%); for familiar word reading, scores increased from 1.48 words per minute to 23.15 words per minute (a change of over 1,400%); for non-word reading, scores increased from 1.03 correctly decoded non-words per minute at baseline to 18.55 correctly decoded non-words at the end-of-project survey (a change of almost 1,700%) and finally, for oral reading fluency, scores increased from 3.49 correctly decoded words per minute at the baseline to 28.39 correctly decoded words at the end-of-project (a change of over 700%). In the following, the EGRA subtasks results will be presented in detail, with a special focus on gender differences. Note that the data allows only to describe how students' skills changed between baseline and end-of-project assessment; it is not possible to establish the causal contribution of the SENSE program.

				Mean	Std.Err.	95% Conf. Interval	
clspm	**Letter sounds**	**Baseline**		4.27	0.317	3.65	4.89
		End-of-project		39.71	0.804	38.13	41.28
		Change	**absolute**	35.44***			
			relative	829%			
			F-statistics	F(1, 1673) = 1733.26			
			p-value	0.000			
csspm	**Syllable sounds**	**Baseline**		2.47	0.305	1.87	3.07
		End-of-project		36.15	0.869	34.45	37.86
		Change	**absolute**	33.68***			
			relative	1362%			
			F-statistics	F(1, 1674) = 1385.12			
			p-value	0.000			
cwpm	**Familiar word reading**	**Baseline**		1.48	0.200	1.09	1.87
		End-of-project		23.15	0.671	21.83	24.46
		Change	**absolute**	21.67***			
			relative	1463%			
			F-statistics	F(1, 1681) = 970.63			
			p-value	0.000			
cnonwpm	**Nonword reading**	**Baseline**		1.03	0.161	0.72	1.35
		End-of-project		18.55	0.582	17.41	19.69
		Change	**absolute**	17.52***			
			relative	1696%			
			F-statistics	F(1, 1678) = 846.06			
			p-value	0.000			
orf	**Oral reading fluency**	**Baseline**		3.49	0.469	2.569	4.41
		End-of-project		28.39	0.781	26.86	29.92
		Change	**absolute**	24.90***			
			relative	713%			
			F-statistics	F(1, 1651) = 770.72			
			p-value	0.000			

Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

Table 3 Overview of results of timed tasks, by survey round

4.1. PERCENTAGE SCORING ZERO

Table 4 gives an overview of the percentage of students scoring zero on each subtask in each survey round. The percentage of zero-scores is generally high at the baseline across subtasks, but decreased dramatically by the end of the project. These changes are highly statistically significant. As expected, both at baseline and at the end-of-project assessment students are more likely to score zero on the more advanced tasks. In The foundational task letter sound identification, about 48% of students scored at the baseline; this number decreased by 45%-points to only 2.5% at the end-of-project assessment. For syllable sound identification, the percentage of zero scores decreased from 81% to under 8%; for familiar word reading, from 90% to 16%; and for non-word reading, from 92% to 22%; and for oral reading fluency, from 88% to 15.7%. For both reading comprehension and listening comprehension, two statistics are reported: first, the percentage of students scoring zero among all participants, and second, the percentage of students scoring zero among those who attempted the task (for instance, only students who managed to actually read a text went on to attempt to answer reading comprehension questions). Hence, while the overall percentage of students scoring zero in the reading comprehension task decreased dramatically from 84% to 23%, the respective percentage among those who actually attempted the task slightly increased, from 5.2% to 8%. This can be explained by the fact that more students managed to read the text, but then struggled to answer the questions; it should thus not be interpreted as a decrease in reading skills. Figure 2 gives a visual overview of these results.

In the remainder of the report, all statistics are reported including students scoring zero, unless explicitly stated otherwise.

	Baseline				End-of-project				Change	
	Mean	SE	95% Conf. Interval		Mean	SE	95% Conf. Interval		Abs.	%
clspm	47.9%	1.58%	44.8%	51.0%	2.5%	0.60%	1.4%	3.7%	-45.4%***	-94.7%
csspm	84.1%	1.21%	81.7%	86.5%	7.8%	0.95%	5.9%	9.7%	-76.3%***	-90.7%
cwpm	90.2%	1.01%	88.2%	92.2%	16.3%	1.26%	13.9%	18.8%	-73.9%***	-81.9%
cnonwpm	92.1%	0.92%	90.3%	93.9%	21.6%	1.35%	18.9%	24.2%	-70.5%***	-76.5%
orf	87.9%	1.07%	85.8%	90.0%	15.7%	1.19%	13.4%	18.0%	-72.2%***	-82.1%
rcomp (all)	83.9%	1.11%	81.7%	86.1%	22.8%	1.30%	20.2%	25.3%	-61.1%***	-72.8%
rcomp (nm.)	5.2%	0.73%	3.8%	6.7%	8.0%	0.89%	6.3%	9.8%	2.8%*	53.4%
lcomp (all)	13.2%	1.05%	11.2%	15.3%	1.0%	0.27%	0.4%	1.5%	-12.2%***	-92.8%
lcomp (nm.)	7.1%	0.84%	5.5%	8.7%	0.7%	0.23%	0.2%	1.1%	-6.4%***	-90.5%

Key: clspm = letter sounds; sccpm = syllable sounds; cwpm = familiar words; cnonwpm = invented words; rcomp = reading comprehension; lcomp = listening comprehension; nm = "non-missing only"
Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

Table 4 Percentage of students scoring zero on each subtask, by survey round and state

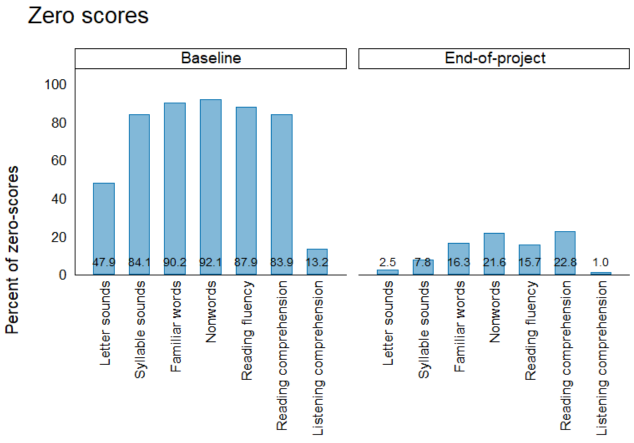


Figure 2 Percentage of students scoring zero, by subtask and survey round

4.2. LETTER SOUND IDENTIFICATION

As Table 5 as well as Figure 3 show, the proficiency level in the foundational skill of letter sound increased over the course of the project, for both boys and girls. At the baseline, students were able to identify an average of 4.27 letter sounds per minute correctly; by the end-of-project assessment, this increased to 39.7. Boys improved from 4.5 to 39.4 correct letter sounds per minute; girls, from 4.0 to 40.0. All of these improvements are highly statistically significant, but the differences between boys and girls are not. The right side of Table 5 shows the assessment results by skill level category. While at the baseline, over 90% of students could identify less than 20 letter sounds correctly, at the end of the project over 40% of learners could correctly identify more than 40 letter sounds correctly (almost 19% even fell into the top-performing group of 60 or more correct letter sounds per minute). These positive changes are statistically significant, and can be observed for both girls and boys.

		Average correct letter sounds per minute (clspm)				Percentage per category of correct letter sounds per minute (clspm)				
		Mean	**SE**	**95% Conf. Int.**		**0**	**1-19,9**	**20-39,9**	**40-59,9**	**60+**
Male	**Baseline**	4.53	0.553	3.45	5.62	50.70%	42.80%	4.62%	1.06%	0.83%
	End-of-project	39.37	1.210	37.00	41.74	2.43%	16.90%	40.30%	22.20%	18.30%
	Difference	34.84***								
	F-Statistic	F(1, 1635) = 692.77				F(3.96, 6495.40)= 123.2671				
	p-value	0.000				0.000				
Female	**Baseline**	4.04	0.378	3.29	4.78	45.30%	49.10%	3.96%	0.89%	0.71%
	End-of-project	40.05	1.293	37.51	42.58	2.60%	19.00%	35.90%	23.00%	19.60%
	Difference	36.01***								
	F-Statistic	F(1, 1665) = 716.66				F(3.91, 6522.71)= 127.9				
	p-value	0.000				0.000				
Total	**Baseline**	4.27	0.317	3.65	4.89	47.90%	46.10%	4.27%	0.97%	0.77%
	End-of-project	39.71	0.804	38.13	41.28	2.52%	17.90%	38.10%	22.60%	18.90%
	Difference	35.44***								
	F-Statistic	F(1, 1673) = 1733.26				F(3.89, 6539.73)= 248.8				
	p-value	0.000				0.000				

Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

Table 5 Letter sound identification: results by round and gender

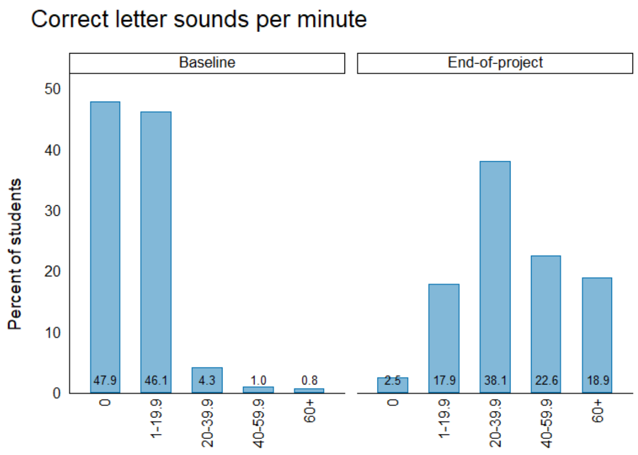


Figure 3 Distribution of letter sound scores (clspm), by survey round

4.3. SYLLABLE IDENTIFICATION

A large performance increase can also be observed for syllable identification, as summarized in Table 6 and Figure 4. The average score on this subtask increased from 2.47 correct syllable sounds at the baseline to 36.15 at the end-of-project, with similar changes happening for boys and girls. Again, all these differences are statistically significant, while the difference between boys and girls is not. At the baseline, over three quarters of learners had fallen into the worst performance category, and a further eighth could only identify between 1 and 19.9 syllable sounds per minute correctly; after a year in the project, only 2.3% could not identify any syllable sound, and an eighth could only identify between 1 and 19.9 syllable sounds – while approximately half of students fell into the categories 20-39.9 and 40 to 59.9 correct syllable sounds, and 30.6% even fell into the highest category of 60+ correct syllable sounds.

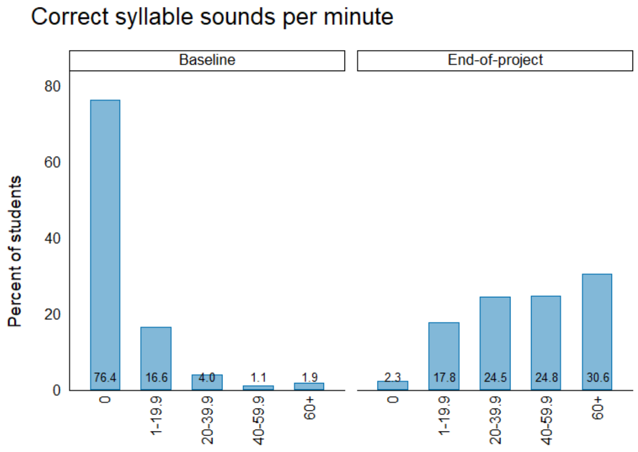


Figure 4 Distribution of syllable sound scores (csspm), by survey round

		Average correct syllable sounds per minute (csspm)				Percentage per category of correct syllable sounds per minute (csspm)				
		Mean	**SE**	**95% Conf. Int.**		**0**	**1-19,9**	**20-39,9**	**40-59,9**	**60+**
Male	**Baseline**	3.11	0.509	2.11	4.11	77.20%	13.20%	6.47%	1.01%	2.12%
	End-of-project	35.62	1.282	33.10	38.13	2.19%	15.60%	26.70%	26.40%	29.10%
	Difference	32.51***								
	F-Statistic	F(1, 1637) = 561.41				F(3.90, 5032.48)= 94.8000				
	p-value	0.000				0.000				
Female	**Baseline**	1.89	0.354	1.20	2.59	75.50%	20.00%	1.60%	1.21%	1.63%
	End-of-project	36.69	1.418	33.91	39.47	2.43%	19.90%	22.40%	23.30%	32.00%
	Difference	34.80***								
	F-Statistic	F(1, 1664) = 568.08				F(3.98, 5099.83)= 104.3514				
	p-value	0.000				0.000				
Total	**Baseline**	2.47	0.305	1.87	3.07	76.40%	16.60%	4.03%	1.11%	1.87%
	End-of-project	36.15	0.869	34.45	37.86	2.31%	17.80%	24.50%	24.80%	30.60%
	Difference	33.68***								
	F-Statistic	F(1, 1674) = 1385.12				F(3.99, 4108.82)= 184.0674				
	p-value	0.000				0.000				

Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

Table 6 Syllable sound identification: results by round and gender

4.4. FAMILIAR WORD READING

The positive trend continues for the subtask familiar word reading (see Table 7). On average, students could read only 1.48 familiar words per minute at the baseline, but they improved to 23.15 words at the end of the project. Girls and boys both improved at a similar pace, with again no differences being detectable among genders. Note that despite this progress, the average results are still below the benchmark of 55 correct words per minute (see section 5.1.5). However, the percentage of students being able to read 60 or more words per minute octuplicated over the course of the project, from 0.26% at the baseline to 1.98% at the end-of-project assessment. Similarly, the share of students not being able to read a single word decreased dramatically from over 90% to around 16%. At the end of the project, the relative majority of students (32.3%) can be found in the performance category of 1-19.9 correct words per minute, followed by the category 20-39.9 correct words per minute (29.8% of students). Figure 5 summarizes these results graphically.

		Average correct familiar words per minute (cwpm)				Percentage per category of correct familiar words per minute (cwpm)				
		Mean	SE	95% Conf. Int.		0	1-19,9	20-39,9	40-59,9	60+
Male	**Baseline**	1.78	0.325	1.14	2.42	89.20%	7.43%	2.40%	0.59%	0.40%
	End-of-project	23.44	1.033	21.41	25.47	15.20%	32.60%	31.60%	18.60%	1.98%
	Difference	21.66***								
	F-Statistic	F(1, 1639) = 401.01				F(3.91, 6416.39)= 120.5307				
	p-value	0.000				0.000				
Female	**Baseline**	1.21	0.240	0.74	1.68	91.10%	7.10%	1.15%	0.53%	0.14%
	End-of-project	22.85	0.981	20.93	24.77	17.50%	32.10%	28.00%	20.50%	1.97%
	Difference	21.64***								
	F-Statistic	F(1, 1669) = 458.94				F(3.93, 6563.41)= 141.5368				
	p-value	0.000				0.000				
Total	**Baseline**	1.48	0.200	1.09	1.87	90.20%	7.26%	1.75%	0.56%	0.26%
	End-of-project	23.15	0.671	21.83	24.46	16.40%	32.30%	29.80%	19.60%	1.98%
	Difference	21.67***								
	F-Statistic	F(1, 1681) = 970.63				F(3.95, 6636.51)= 257.1841				
	p-value	0.000				0.000				

Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

Table 7 Familiar word reading: results by round and gender

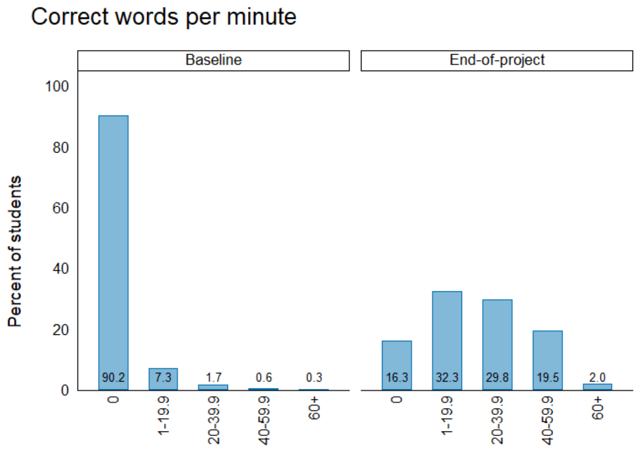


Figure 5 Distribution of familiar word reading scores (cwpm), by survey round

4.5. NON-WORD DECODING

Non-word decoding is an essential skill, as it shows the ease with which learners can read unfamiliar words. It is expected that students score lower on this task than on the subtask of familiar word reading, which is indeed the case, though the differences are small. Table 8 as well as Figure 6 present the results of the comparison between survey rounds. Overall, learners correctly decoded 1.03 non-words per minute at the baseline; after a year in the project, learners were able to correctly decode 18.55 non-words per minute. This average lies comfortably above the DIBELS benchmark of 13 correct non-words per minute (see section 5.1.5) Again, improvement rates of boys and girls are comparable and not distinguishable statistically. A look at the results by performance categories (right part of Table 8) shows a similar picture as for the subtask familiar word reading: while at the baseline, 92% of learners could not correctly decode a single invented word, the same was only true for 21.6% of learners at the end of the project. At the same time, the share of students falling into the highest two performance categories increased from 0.4% to 12%.

		Average correct invented words per minute (cnonwpm)				Percentage per category of correct invented words per minute (cnonwpm)				
		Mean	**SE**	**95% Conf. Int.**		**0**	**1-19,9**	**20-39,9**	**40-59,9**	**60+**
Male	**Baseline**	1.25	0.247	0.77	1.73	91.60%	7.09%	0.68%	0.38%	0.21%
	End-of-project	18.70	0.881	16.97	20.43	20.40%	38.40%	28.70%	11.00%	1.57%
	Difference	17.45***								
	F-Statistic	F(1, 1638) = 364.19				F(3.82, 6266.10)= 124.6196				
	p-value	0.000				0.000				
Female	**Baseline**	0.84	0.216	0.41	1.26	92.50%	6.21%	1.05%	0.05%	0.14%
	End-of-project	18.39	0.896	16.63	20.15	22.90%	35.00%	30.40%	10.10%	1.49%
	Difference	17.55***								
	F-Statistic	F(1, 1667) = 361.22				F(3.41, 5672.80)= 157.4549				
	p-value	0.000				0.000				
Total	**Baseline**	1.03	0.161	0.72	1.35	92.10%	6.63%	0.87%	0.21%	0.18%
	End-of-project	18.55	0.582	17.41	19.69	21.60%	36.70%	29.60%	10.50%	1.53%
	Difference	17.52***								
	F-Statistic	F(1, 1678) = 846.06				F(3.88, 6513.92)= 268.8902				
	p-value	0.000				0.000				

Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

Table 8 Non-word decoding: results by round and gender

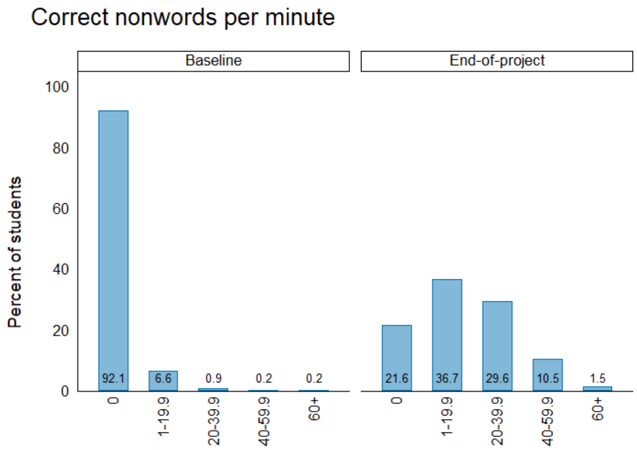


Figure 6 Distribution of nonword reading scores (cnonwpm), by survey round

4.6. ORAL READING FLUENCY

Oral reading fluency is perhaps the central task on the EGRA assessment, as it measures the ease with each a student can read a connected text. Table 9 as well as Figure 7 present the results of the comparison between survey rounds. Overall, learners correctly read 3.49 words per minute at the baseline; at the end-of-project assessment, learners were able to correctly read 28.39 words per minute. Again, improvement rates of boys and girls are comparable and not distinguishable statistically. The results by performance categories (right part of Table 9) are comparable to those for the subtasks familiar word reading and non-word reading: while at the baseline, 88% of learners could not correctly read a single word, the same was only true for 15.7% of learners at the end of the project. At the same time, the share of students falling into the highest performance category increased from 3% to 12.7%.

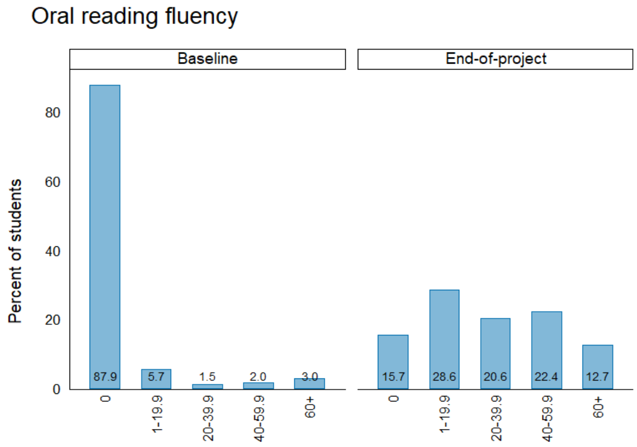


Figure 7 Distribution of oral reading fluency scores (orf), by survey round

		Average correct words per minute (orf)				Percentage per category of correct words per minute (orf)				
		Mean	SE	95% Conf. Int.		0	1-19,9	20-39,9	40-59,9	60+
Male	Baseline	4.29	0.727	2.86	5.72	84.90%	7.31%	2.19%	2.94%	2.64%
	End-of-project	28.92	1.222	26.52	31.32	14.70%	28.90%	20.90%	22.30%	13.30%
	Difference	24.63***								
	F-Statistic	F(1, 1626) = 300.64				F(3.94, 6449.63)= 108.3498				
	p-value	0.000				0.000				
Female	Baseline	2.762	0.605	1.575	3.949	90.50%	4.22%	0.82%	1.11%	3.32%
	End-of-project	27.85	1.173	25.55	30.15	16.70%	28.40%	20.30%	22.50%	12.10%
	Difference	25.09***								
	F-Statistic	F(1, 1649) = 364.32				F(3.90, 6515.43)= 143.3920				
	p-value	0.000				0.000				
Total	Baseline	3.49	0.469	2.569	4.41	87.90%	5.69%	1.47%	1.98%	2.99%
	End-of-project	28.39	0.781	26.86	29.92	15.70%	28.60%	20.60%	22.40%	12.70%
	Difference	24.90***								
	F-Statistic	F(1, 1651) = 770.72				F(3.91, 6571.86)= 246.8776				
	p-value	0.000				0.000				

Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

Table 9 Oral reading fluency: results by round and gender

4.7. ORAL READING COMPREHENSION

The ultimate objective is for students to be able to retrieve information from a text they read. The subtask of oral reading comprehension is thus of particular relevance, as it measures the number of comprehension questions a student is able to answer after reading a short grade-level appropriate narrative. Results for this sub-skill show, in line with the previous results, an important and statistically significant increase in proficiency over the duration of the project. As summarized in Table 10 and in Figure 8, on average, learners at the baseline were able to correctly answer 0.53 out of five comprehension questions, while this number increased by project end to 3.10. At the baseline, boys were slightly more proficient than girls in this particular subtask, yet both improved by a lot, and at the end of the project no difference between genders could be established anymore. While at the baseline, the vast majority of students (84%) were not able to correctly answer a single comprehension question, the same was true and the end-of-project assessment for only 23% of students. Conversely, while at the baseline only 4.5% of students were able to correctly answer all five comprehension questions, 42.1% of students were able to do so after a year in the project.

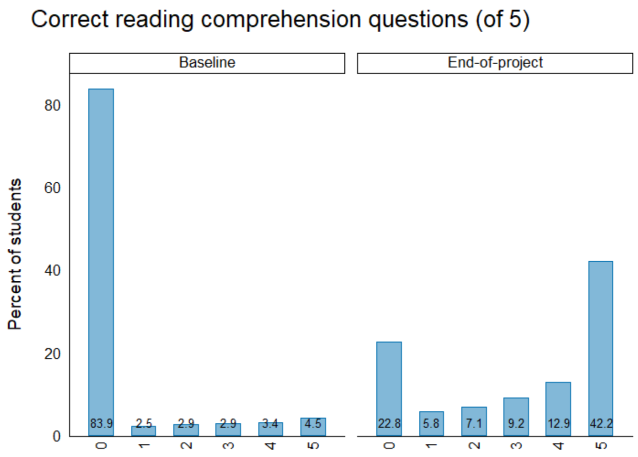


Figure 8 Distribution of reading comprehension scores, by survey round

		Number of correctly answered questions (oral comprehension)				Percentage per number of correctly answered questions (oral comprehension)					
		Mean	**SE**	**95% Conf. Int.**		**0**	**1**	**2**	**3**	**4**	**5**
Male	**Baseline**	0.59	0.067	0.46	0.72	82.6%	2.4%	3.1%	2.9%	3.1%	5.9%
	End-of-project	3.11	0.098	2.92	3.30	23.3%	4.8%	7.2%	8.8%	13.8%	42.1%
	Difference	2.52***									
	F-Statistic	F(1, 1640) = 469.90				F(4.94, 8102.67)= 59.1923					
	p-value	0.000				0.000					
Female	**Baseline**	0.47	0.055	0.36	0.58	85.0%	2.5%	2.7%	3.0%	3.6%	3.2%
	End-of-project	3.09	0.098	2.90	3.28	22.2%	6.8%	7.0%	9.7%	12.1%	42.2%
	Difference	2.62***									
	F-Statistic	F(1, 1669) = 553.41				F(4.95, 8266.03)= 74.5841					
	p-value	0.000				0.000					
Total	**Baseline**	0.53	0.041	0.45	0.61	83.9%	2.5%	2.9%	2.9%	3.4%	4.5%
	End-of-project	3.10	0.057	2.99	3.21	22.8%	5.8%	7.1%	9.2%	12.9%	42.1%
	Difference	2.57***									
	F-Statistic	F(1, 1682) = 1416.67				F(4.90, 8234.87)= 135.7605					
	p-value	0.000				0.000					

Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

Table 10 Oral comprehension: results by round and gender

4.8. LISTENING COMPREHENSION

The final subtask that learners were asked to perform was listening comprehension: After a short, grade-level appropriate text was read out to them, students were asked five comprehension questions, which are designed to test their comprehension ability without the added challenge of having to decode a text. It is expected that students perform better on this task than on the oral reading comprehension task, which is in fact the case. Additionally, the EGRA results show an improvement in this skill, albeit less dramatic than in other subtasks – because results had already been better at the baseline. On average, students at the baseline had correctly answered three out of the five comprehension questions; by project end, this number increased to 4.4. Results for boys and girls are similar and not statistically distinguishable. Contrary to previously described tasks, the largest share of learners had already fallen into the highest performance categories at the baseline, with only 13.2% of students not being able to correctly answer a single question, and almost half of students (46.5%) being able to answer four or five of the five questions correctly. By contrast, at the end-of-project assessment, only 1% of students fell into the lowest category, while 85% of students could correctly answer four or five questions. The results are summarized in Table 11 as well as in Figure 9.

		Number of correctly answered questions (listening comprehension)				**Percentage per number of correctly answered questions (listening comprehension)**					
		Mean	**SE**	**95% Conf. Int.**		**0**	**1**	**2**	**3**	**4**	**5**
Male	**Baseline**	3.06	0.075	2.91	3.21	11.5%	8.3%	14.6%	18.6%	22.1%	24.8%
	End-of-project	4.43	0.046	4.34	4.52	1.0%	1.1%	2.8%	10.0%	19.7%	65.5%
	Difference	1.37***									
	F-Statistic	F(1, 1640) = 236.26				F(4.89, 8016.81)= 38.1507					
	p-value	0.000				0.000					
Female	**Baseline**	2.94	0.077	2.79	3.09	14.8%	8.1%	14.7%	16.2%	22.7%	23.4%
	End-of-project	4.36	0.051	4.26	4.46	0.9%	3.2%	2.9%	8.1%	21.7%	63.2%
	Difference	1.42***									
	F-Statistic	F(1, 1669) = 235.52				F(4.87, 8134.03)= 40.8850					
	p-value	0.000				0.000					
Total	**Baseline**	3.00	0.051	2.90	3.10	13.2%	8.2%	14.7%	17.4%	22.4%	24.1%
	End-of-project	4.39	0.032	4.33	4.46	1.0%	2.1%	2.8%	9.1%	20.7%	64.3%
	Difference	1.40***									
	F-Statistic	F(1, 1682) = 519.26				F(4.85, 8161.82)= 81.5483					
	p-value	0.000				0.000					

Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

Table 11 Listening comprehension: results by round and gender

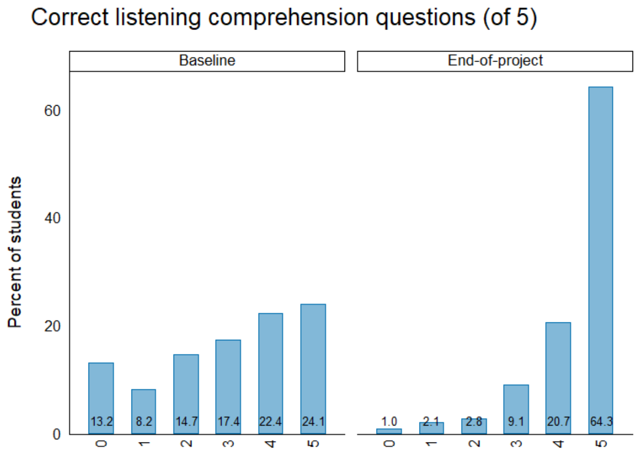


Figure 9 Distribution of listening comprehension scores, by survey round

5. GLOBAL PROFICIENCY FRAMEWORK

5.1. MAPPING EGRA SCORES TO THE GLOBAL PROFICIENCY FRAMEWORK

In 2019, USAID in collaboration with a range of organizations published the *Global Proficiency Framework (GPF) for Reading and Mathematics, Grades 2 to 6.* This document defines the proficiency expectations for reading and mathematics for students throughout primary education (grades 2 through 6), and thus aims to facilitate reporting for the Sustainable Development Goals 4.1.1(a) and (b): Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex (USAID et al. 2019). The GPF represents significant process in international education evaluation in that it provides a detailed description of different skills and proficiency levels that students are expected to achieve, allowing for benchmarking at all levels of the education system. However, being only a framework, the specific ways in which the proficiency levels are to be measured remain to be decided by individual actors.

As part of the SENSE project, AUN is using the Early Grade Reading Assessment (EGRA) (RTI International 2009; 2016) to assess the Hausa literacy level of its beneficiaries. The EGRA tool itself does not specify expected proficiency levels, which raises the question how results on reading proficiency, as measured through the EGRA tool, can be interpreted in the light of the new GPF framework. While the EGRA tool was developed as a diagnostic tool and not designed for international (or even national inter-language) comparisons, it is worth exploring to what extent its results can be translated into proficiency levels defined by the GPF. When the SENSE project started developing a mapping system to interpret EGRA results in the GPF framework, there was no published system to map Hausa EGRA results onto the PGF framework. Therefore a mapping system was developed for the purposes of the SENSE project.. After this mapping exercise for the SENSE project was completed, MSI, in collaboration with US Agency for International Development (USAID), the Nigerian Federal Ministry of Education (FME), the Universal Basic Education Commission (UBEC), and the Nigerian Educational Research and Development Council (NERDC) published a draft report describing efforts to set global benchmarks for the Hausa EGRA in order to link it with the GPF (Management Systems International (MSI) 2020). However, due to the use of different EGRA assessment tools, the benchmarks developed by MSI are not applicable to the SENSE project, which why the SENSE project continues to use the mapping system developed for this project.

The remainder of this section gives an overview of the GPF framework; reviews the proficiency scores available in the EGRA tool used; and describes the EGRA-GPF mapping applied in the SENSE project, including a discussion on the applicability of MSI's preliminary mapping proposal as an alternative for the SENSE project. Key aspects are as follows:

- In order to meet grade-level proficiency in listening comprehension, a student needed to answer correctly to at least three out of the five listening comprehension questions
- In order to meet grade-level proficiency in decoding, a student needed to obtain an orf score of at least 30 words per minute
- In order to meet grade-level proficiency in reading comprehension, a student needed to answer correctly to at least three out of the five reading comprehension questions
- In order to meet overall grade-level proficiency, a student needed to at least achieve minimum grade-level proficiency in all three subdomains.

5.1.1. THE GPF PROFICIENCY LEVELS AND GLOBAL PROFICIENCY DESCRIPTORS

The GPF defines four proficiency levels (USAID et al. 2019, 2):

DOES NOT MEET MINIMUM PROFICIENCY

Learners lack the most basic knowledge and skills. As a result, they generally cannot complete the most basic task.

PARTIALLY MEETS MINIMUM PROFICIENCY

Learners have limited knowledge and skills. As a result, they can partially complete basic tasks.

MEETS MINIMUM PROFICIENCY

Learners have developed sufficient knowledge and skills. As a result, they can successfully complete basic tasks.

EXCEEDS MINIMUM PROFICIENCY

Learners have developed superior knowledge and skills. As a result, they can successfully complete complex tasks.

Furthermore, the GPF defines expected proficiency for grade 2 learners in three age-relevant literacy domains as follows (USAID et al. 2019, 30–31):

Aural Language Comprehension

Given a text that is read to them learners can:

- Identify key events, ideas or major characters
- Make simple inferences
- Identify the meaning of key words

Decoding

Given a short grade-level text learners can:

- Decode most words in a connected text, including some unfamiliar ones

Reading comprehension/Retrieving Information

Given a grade-level narrative or expository text, learners can:

- Identify the meaning of most unfamiliar words or familiar words used in unfamiliar ways (i.e., homophones)
- Locate most pieces of explicit information in a sentence when the information is prominent and there is no competing information.

Finally, the framework goes on to describe performance expectations for four defined proficiency levels, as summarized in the table below.

Does not meet minimum proficiency	**Partially meets minimum proficiency**	**Meets minimum proficiency**	**Exceeds minimum proficiency**
Aural language comprehension			
Retrieve and interpret information at sentence/text level: Identify explicit and implicit information in text read to learner			
Performance is below partially meets minimum proficiency	Identify simple inferences within single sentences	Identify simple inferences across consecutive sentences	Identify simple inferences by connecting information across the text
Retrieve information at word level: Understand the meaning of words in a text read to the learner			
Performance is below partially meets minimum proficiency	When listening to longer texts, identify the meaning of very familiar words	When listening to longer texts, identify the meaning of familiar words and some unfamiliar words	When listening to longer texts, identify the meaning of familiar and unfamiliar words
Decoding			
Precision – decode accurately a short, grade-level connected text			
Performance is below partially meets minimum proficiency.	Decode very familiar words in connected text accurately; makes frequent errors	Decode familiar words in connected text accurately, but reads slowly, word by word	Decode words in connected text accurately, including unfamiliar words, at a pace that supports basic understanding
Reading comprehension of simple, grade 2-level connected text			
Retrieve information at word level: Understand in connected text the meaning of unfamiliar words, or of familiar words used in unfamiliar ways (i.e., homophones)			
Performance is below partially meets minimum proficiency	Identify the meaning of very familiar words but has difficulty identifying the meaning of familiar words when they have	Identify the meaning of familiar words, including when they have regular morphological changes	Identify the meaning of familiar and unfamiliar words

	regular morphological changes		
Retrieve information at sentence or text level: Retrieve prominent information when information is found in a single sentence containing no competing information. The information is generally a response to a 'who, what, when and where" question and the information sought is generally names, facts or numbers.			
Performance is below partially meets minimum proficiency	Retrieve explicit pieces of information by direct word matching (e.g., answers the question, 'What is the girl's name?' when the text says, 'The girl's name is Dana.'	Retrieve explicit pieces of information from a single sentence	Retrieve explicit pieces of information across more than one sentence

Table 12 Global proficiency descriptors. Reproduced from USAID et al. (2019, 40–41)

5.1.2. EGRA PROFICIENCY SCORES

EGRA was developed to measure student literacy learning in early grades in low-income countries, as no simple, effective, and low-cost measure for student learning had been available for this context (RTI International 2009). Today there are several versions of the EGRA tool (RTI International 2016); for the project at hand, the NEI+ Mukaranta EGRA instrument was used.

Each EGRA survey consists of a set up subtasks that aim to measure different components of literacy learning. The EGRA tools used for this project contain the following subtasks (with the proficiency indicator in parenthesis):

- Letter identification (correct letter sounds per minute, clspm)
- Syllable identification (correct syllable sounds per minute, csspm)
- Familiar word reading (correct words per minute, cwpm)
- Non-word reading (correct non-words per minute, cnonwpm)
- Oral reading fluency (correct words per minute in a connected text, orf[1])
- Oral reading comprehension
- Listening comprehension

For each subtasks, proficiency scores are calculated separately; there is no overall EGRA score or proficiency classification. Instead, the authors of the EGRA toolkit suggest that national benchmarks be developed (RTI International 2016). No such official national benchmarks are available for Nigeria.

5.1.3. A PROPOSED EGRA/GPF MAPPING

5.1.3.1. SENSE PROJECT MAPPING

The GPF defines grade-level proficiency is across three domains and five subdomains, as described above. Out of the seven EGRA subtasks, only the three (oral reading fluency, oral reading comprehension, and listening [aural] comprehension) are present in the GPF. The GPF is, however, much more specific about different literacy

[1] The abbreviation orf-cwpm is used in this section to make a distinction to cwmp while making clear that this is also a words-per-minute measure.

subdomains than what EGRA allows to assess. There is thus no obvious mapping. For aural language comprehension, while the EGRA tool only allows to count the number of correctly answered comprehension questions, as shown in Table 12 the GPF defines to what extent students should be able to retrieve and interpret information, and to what extent they are able to understand the meaning of words. For oral reading fluency, EGRA data allows to determine the number of correctly decoded words, total or per minute, without providing a benchmark of what should be considered “slow” or “fast” reading, while the GPF makes a qualitative difference between familiar and unfamiliar words and by reading speed. For reading comprehension, the EGRA tool used for this project only allows to count the number of correctly answered comprehension questions, while as shown in Table 12 the GPF asks to what extent a student is able to understand the meaning of familiar and unfamiliar words, and retrieve information from the text.

Given these factors, it is not possible to assess the proficiency level of all GPF subdomains using the EGRA data. Instead, the SENSE project uses the EGRA subtasks listening comprehension, oral reading fluency, and reading comprehension are to determine GPF proficiency levels for aural comprehension, decoding and reading comprehension, respectively. Overall proficiency is then defined as the *lowest* level of proficiency that a student achieves across these subdomains, as a student can only be considered to meet the grade level minimum proficiency if he or she performs adequately in all subdomains. The SENSE project thus uses a two-step approach to define grade-level proficiency: first, a student’s level of proficiency is determined for each of the three GPF subdomains for which EGRA assessment data is available; and second, a students’ overall level of proficiency is determined as lowest of these three subdomain proficiency levels. For instance, if a student *Meets minimum proficiency* in aural comprehension and decoding, but only *Partially meets minimum proficiency* in reading comprehension, the student’s level of grade-level proficiency is defined as *Partially meets minimum proficiency*. This is a conservative approach that assures that improvements in comparatively easier subtasks do not artificially inflate overall improvement rates.

After briefly comparing this approach to the mapping (“policy linking”) proposed by MSI, the following subsections describe how the SENSE project uses EGRA data to determine a student’s level of proficiency for each GPF subdomain.

5.1.3.2. COMPARISON WITH THE PROPOSED MSI MAPPING FOR NIGERIA (“POLICY LINKING”)

The mapping (“policy linking”) exercise carried out by MSI for Nigeria has been an extensive, multi-stakeholder process based on an established methodology, aiming to lend credibility and validity to the benchmarking. The mapping was carried out in three separate tasks (Management Systems International (MSI) 2020): First, a panel of experts determined whether and to what extent the EGRA tool was suitable for linking to GPF (it was); second, the panel assessed how the assessment items from the EGRA tool matched the performance standards (proficiency descriptors) of the GPF, resulting in an assignment of each item to a specific performance standard; and third, the panelists identified global benchmarks based on the Angoff method. This last step involved field-testing each item in the classroom to assess its difficulty, by testing what type of student could correctly answer the given item. The results of this test were then used to construct benchmarks, which represented the number of items that two out of three students of each student type could answer correctly. Note that this linking approach does not attempt to define proficiency levels for the individual GPF domains (or subdomains), but rather constructs an overall proficiency benchmark across all items and domains, respectively.

The MSI panel concluded that only two EGRA subtasks that formed part of the NEI+ EGRA assessment tool can be used for the benchmarking exercise and be mapped onto (linked to) the GPF: “oral reading fluency” and “oral reading comprehension". The NEI+ Mukaranta EGRA applied by SENSE includes additionally the subtask of listening comprehension, allowing to also assess the GPF domain Aural Language Comprehension.

Importantly, the benchmarks resulting from the MSI mapping process are specific to the precise EGRA instrument used, as different versions of the EGRA tool use different word lists and/or reading passages and thus vary slightly in the number of test items as well as in the global proficiency standard necessary to answer any given item. For instance, one EGRA tool might include 15 out of 40 items that two out of three students who typically meet grade level standards answer correctly, while another EGRA tool might include 20 out of 50 such items. Hence, unless the same EGRA tool is used, the benchmarks cannot be directly transferred. The MSI mapping uses the NEI+ 2018 midline EGRA with an oral reading passage containing 35 items (words), while SENSE uses a version of the NEI+ EGRA with an oral reading passage containing 60 items (words). Hence, the benchmarks are not directly applicable for this project, and the SENSE project continues to use GPF mapping approach developed for this project.

5.1.4. AURAL LANGUAGE COMPREHENSION (LISTENING COMPREHENSION)

5.1.4.1. RETRIEVE AND INTERPRET INFORMATION AT SENTENCE/TEXT LEVEL: IDENTIFY EXPLICIT AND IMPLICIT INFORMATION IN TEXT READ TO LEARNER

The Hausa EGRA listening comprehension task applied for his assessment consists of a short text read to students, and five comprehension questions. In order to assess the GPF proficiency level, each of the five comprehension questions can be assigned to one of the proficiency level definitions outlined in the GPF (see annex II for details). Table 13 shows how the EGRA results will be interpreted in the GPF for this subdomain.

Proficiency level	GPF Proficiency level definition	EGRA correspondence
Exceeds minimum proficiency	Identify simple inferences by connecting information across the text	Student responds correctly to all level 3 questions and at least two other questions, OR Student responds correctly to all five questions
Meets minimum proficiency	Identify simple inferences across consecutive sentences	Student responds correctly to the level-2 question (one questions) and all level 1 questions (two questions) OR Student responds correctly to at any three questions*
Partially meets minimum proficiency	Identify simple inferences within single sentences	Student responds correctly all level 1-questions (two questions), OR Student responds correctly to any two questions*
Does not meet minimum proficiency	Performance is below partially meets minimum proficiency	Students performing below "partially meets minimum proficiency" requirement

* The key challenge in this mapping approach is that students may, and in fact do, answer higher-level questions correctly but lower-level questions incorrectly. It would be inconsistent to classify a student as not meeting a certain level proficiency, when in fact a higher-level question was answered correctly. Furthermore, as only a total of five questions was asked, there is little room for students to make up for errors. For these reasons, de facto the lowest three proficiency levels are defined simply by the number of questions answered correctly.

Table 13 GPF/EGRA mapping for the literacy subdomain ***Aural language comprehension: Retrieve and interpret information at sentence/text level: Identify explicit and implicit information in text read to learner***

5.1.4.2. RETRIEVE INFORMATION AT WORD LEVEL: UNDERSTAND THE MEANING OF WORDS IN A TEXT READ TO THE LEARNER

The EGRA tool used does not allow to directly assess students' proficiency levels in this subdomain.

5.1.5. DECODING

5.1.5.1. PRECISION – DECODE ACCURATELY A SHORT, GRADE-LEVEL CONNECTED TEXT

EGRA scores can be mapped towards GPF decoding proficiency levels based on the EGRA subtask oral reading fluency[2]. Some assumptions are made on the likelihood of achieving basis reading proficiency based on performance on this subtask.

Official reading fluency benchmarks are not available for Nigeria, or for the Hausa language in general. Other Hausa EGRA Hausa reports, acknowledging that fact, often use widely available English-language benchmarks as a rough guideline (see, for instance, RTI International 2014); some also have set their own benchmarks (Creative Associates International 2018). While English-language benchmarks are well researched, their applicability to Hausa reading is questionable. At the same time, the project-specific benchmarks seem to have been set on an ad-hoc basis, hence their wider applicability is also not clear.

The strategy adopted by the SENSE project is thus to combine the well-research English-language guidelines with experiences from earlier EGRA applications in Hausa to come to some benchmarks. Specifically, the DIBELS reading benchmark for oral reading in English at the beginning of grade 2 is 52 correct words per minute, while students in the range of 37-51 correct words per minute are considered likely to need strategic support. Similarly, the DIBELS benchmark for correct non-words per minute is 13, while those in the range of 6-12 for correct non-words per minute are considered likely to need strategic support (Dynamic Measurement Group Inc 2010).

These benchmarks are roughly consistent with available data on Hausa reading proficiency. The benchmarks for oral reading fluency published on USAID's Early Grade Reading Barometer for Sokoto State (USAID n.d.) and Bauchi State (USAID n.d.) indicate, for specific correct words per minute scores, both the average percent score of reading comprehension for students at the specific oral reading fluency level, and the estimated probability that a student with the specific oral reading fluency score will be able to answer 80% or more of the comprehension questions they are asked about a passage they had just read.

According to this data, grade 2-students reading in Hausa at 50 correct words per minute have, in Sokoto, a 78% chance of understanding at least 80% of a text. Additionally, they have on average a score of 56% on reading comprehension. For Bauchi, the respective numbers differ: grade 2-students reading in Hausa at 50 correct words per minute in Bauchi have a 34% change of understanding at least 80% of a text, and, on average, score 62% on reading comprehension. In order to come close to Sokoto's 78% change of understanding a text, students in Bauchi have to reach 60 correct words per minute. A benchmark between 50 and 60 correct words per minute thus seems to be a reasonable assumption for a "pace that supports basic understanding" (the GPF definition for the highest proficiency level). For this exercise, the average of 55 correct words per minute is thus adapted as the *Exceeds minimum proficiency* benchmark.

[2] The original mapping proposal for the SENSE project used the EGRA subtasks familiar word reading and non-word reading to construct this GPF domain, as the EGRA tool used for the first baseline assessment of the program did not contain an oral reading fluency subtask. This was changed based on feedback from USAID Nigeria that suggested that the EGRA subtask oral reading fluency would be more appropriate for this purpose.

The definition for the GPF's *Meets minimum proficiency* level does not ask for a reading speed that supports basic understanding, but allows students to read slowly, word by word. Considering again the DIBELS range of 37-51 correct words per minute as starting point, a comparison with data from the Early Grade Reading Barometer in Sokoto shows that grade 2-students reading in Hausa at 30 correct words per minute have only a 3% chance of understanding 80% of the text, while the respective value for Bauchi is 30%. This seems acceptable for the GPF definition.[3] Note, however, that this benchmark lies somewhat above a benchmark set by the NEI+ project (Creative Associates International 2018): NEI+ defined 20 correct words per minute for oral reading fluency (and 40% for reading comprehension questions) as the primary 2 benchmark. While these NEI+ benchmarks were set in consultation "with a full range of relevant state and national stakeholders", they are also explicitly meant as "[s]hort-term benchmarks, to be used for the duration of the Initiative" (ibid, 67). Therefore, in light of the discussion above they seem too low to serve as GPF benchmarks. This implies that the benchmark used for the SENSE project is conservative.

Finally, in order to fulfill the *Partially meets minimum proficiency* requirements, students need to be able to decode very familiar words in connected text accurately, even though they make frequent errors. The DIBELS guide is of no help here, yet the evidence from Sokoto and Bauchi suggests that at 10 correct words per minute, the average percentage of reading comprehension falls to 12% and 15%, respectively. At 20 correct words per minute, the respective values are 7% [sic] and 30%. The probability that a student will understand at least 80% of a text lies, for both 10 correct words per minute and 20 correct words per minute, at 0%. Even though text comprehension does not enter in the lower proficiency level of this GPF domain, 20cwpm thus seems to be a reasonable threshold for this lowest proficiency level.

This proficiency mapping is summarized in Table 14.

Proficiency level	GPF Proficiency level definition	EGRA correspondence
Exceeds minimum proficiency	Decode words in connected text accurately, including unfamiliar words, at a pace that supports basic understanding	≥55 orf-cwpm
Meets minimum proficiency	Decode familiar words in connected text accurately, but reads slowly, word by word	30 orf-cwpm – 54 orf-cwpm
Partially meets minimum proficiency	Decode very familiar words in connected text accurately; makes frequent errors	10 orf-cwpm -29 orf-cwpm
Does not meet minimum proficiency	Performance is below partially meets minimum proficiency.	< 10 orf-cwpm

Table 14 GPF/EGRA mapping for the literacy subdomain *Precision – decode accurately a short, grade-level connected text*

[3] This range also include the reading benchmark established at 40 correct words per minute in Malawi for grade 2, in one of the few available EGRA benchmarking exercises in sub-Sahara Africa (Malawi Ministry of Education, Science and Technology 2014)

5.1.6. READING COMPREHENSION OF SIMPLE, GRADE 2-LEVEL CONNECTED TEXT

5.1.6.1. RETRIEVE INFORMATION AT WORD LEVEL: UNDERSTAND IN CONNECTED TEXT THE MEANING OF UNFAMILIAR WORDS, OR OF FAMILIAR WORDS USED IN UNFAMILIAR WAYS (I.E., HOMOPHONES)

The EGRA tool used does not allow to directly assess students' proficiency levels in this subdomain.

5.1.6.2. RETRIEVE INFORMATION AT SENTENCE OR TEXT LEVEL: RETRIEVE PROMINENT INFORMATION WHEN INFORMATION IS FOUND IN A SINGLE SENTENCE CONTAINING NO COMPETING INFORMATION. THE INFORMATION IS GENERALLY A RESPONSE TO A 'WHO, WHAT, WHEN AND WHERE" QUESTION AND THE INFORMATION SOUGHT IS GENERALLY NAMES, FACTS OR NUMBERS.

The NEI+ EGRA reading comprehension tasks consists of a short reading prompt and five comprehension questions. Each of the five comprehension questions can be assigned to one of the proficiency level definitions outlined in the GPF (see annex II for details). The following table shows how the EGRA results will be interpreted in the GPF for this subdomain.

Proficiency level	GPF Proficiency level definition	EGRA correspondence
Exceeds minimum proficiency	Retrieve explicit pieces of information across more than one sentence	Students responds correctly to both level 3 questions, and answers at least four questions correctly
Meets minimum proficiency	Retrieve explicit pieces of information from a single sentence	Student responds correctly to the level-2 and level 1 questions OR Student responds correctly to at any three or four questions*
Partially meets minimum proficiency	Retrieve explicit pieces of information by direct word matching (e.g., answers the question, 'What is the girl's name?' when the text says, 'The girl's name is Dana.	Student responds correctly the level 1-question OR Student responds correctly to any two questions*
Does not meet minimum proficiency	Performance is below partially meets minimum proficiency	Students performing below "partially meets minimum proficiency" requirement

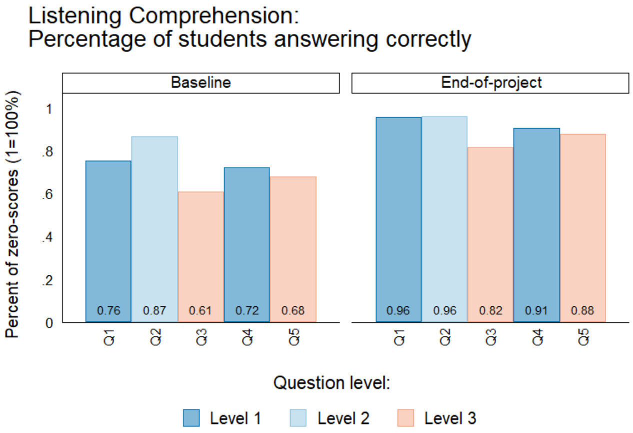
* The key challenge in this mapping approach is that students may, and in fact do, answer higher-level questions correctly but lower-level questions incorrectly. It would be inconsistent to classify a student as not meeting a certain level proficiency, when in fact a higher-level question was answered correctly. Furthermore, as only a total of five questions was asked, there is little room for students to make up for errors. For these reasons, de facto the lowest three proficiency levels are defined simply by the number of questions answered correctly.

Table 15 GPF/EGRA mapping for the literacy subdomain *Reading comprehension: Retrieve information at sentence or text level*

5.2. RESULTS

5.2.1. AURAL LANGUAGE COMPREHENSION (LISTENING COMPREHENSION)

The GPF proficiency level for the subdomain *"Retrieve and interpret information at sentence/text level: Identify explicit and implicit information in text read to learner"* is assessed based on the EGRA listening comprehension subtask. The benchmark for *Meets minimum proficiency* is defined correctly answering to any three (of the five) questions (corresponding to answering correctly all questions corresponding to difficulty/proficiency levels 1 and 2, see annex II). Figure 10 shows the percentage of students answering correctly to any one of the five questions, by the questions' level of difficulty.



Note: All differences between rounds are statistically significant with p<0.001

Figure 10 Figure 11 Percentage of students answering to correctly to any of the five listening comprehension questions, by level of difficulty

Table 16 as well as Figure 12 summarize the results of this mapping, which mirror results on the listening comprehension EGRA subtask presented in section 4.8. At the baseline, the relative majority of learners (33.5%) can be found in the highest proficiency level" Exceeds minimum proficiency", followed by the group in the second highest level, "Meets minimum proficiency (30.2%). 21.5% of did not meet the minimum proficiency level. At the end of the project, only 3.1% of learners did not meet minimum proficiency, while 70% exceed minimum proficiency and 24.1% met minimum proficiency. These improvements are highly statistically significant. Boys and girls improved in similar ways, and no gender differences can be statistically established.

		Does not meet		Partially meets		Meets		Exceeds		Total
		%	95% CI	%	95% CI	%	95% CI	%	95% CI	
Male	Baseline	19.80%	16.46% 23.65%	14.60%	11.69% 18.16%	31.80%	27.55% 36.42%	33.70%	29.52% 38.24%	100%
	End-of-project	2.06%	1.11% 3.78%	2.76%	1.54% 4.90%	24.40%	20.29% 28.95%	70.80%	66.09% 75.14%	100%
	Chi2 Test	F(2.98, 4887.40)=	52.4660							
	p-value	0.000								
Female	Baseline	23.00%	19.46% 26.85%	14.70%	11.87% 18.17%	28.80%	24.95% 33.05%	33.50%	29.46% 37.75%	100%
	End-of-project	4.12%	2.57% 6.54%	2.90%	1.72% 4.84%	23.80%	19.81% 28.32%	69.20%	64.43% 73.56%	100%
	Chi2 Test	F(2.99, 4989.62)=	50.3070							
	p-value	0.000								
All	Baseline	21.50%	19.07% 24.05%	14.70%	12.60% 17.07%	30.20%	27.34% 33.34%	33.60%	30.75% 36.57%	100%
	End-of-project	3.08%	2.13% 4.45%	2.83%	1.99% 4.02%	24.10%	21.25% 27.16%	70.00%	66.87% 72.96%	100%
	Chi2 Test	F(2.97, 4998.26)=	108.7326							
	p-value	0.000								

Table 16 GPF Results Aural Language Comprehension, by survey round and gender

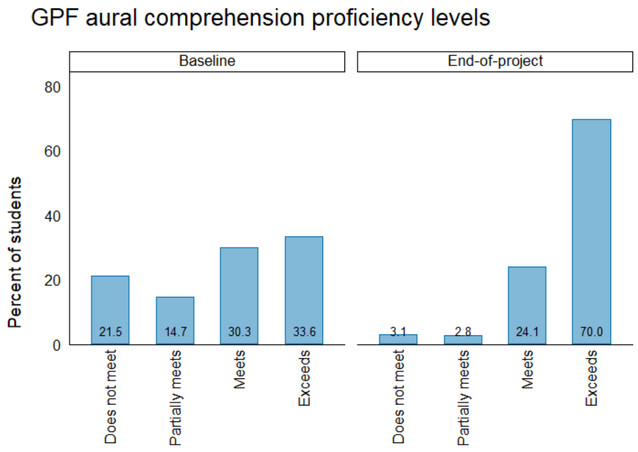


Figure 12 Listening comprehension results in GPF proficiency levels, by survey round

5.2.2. DECODING

The GPF subdomain *"Precision – decode accurately a short, grade-level connected text"* is assessed based on the EGRA subtask oral reading fluency (orf), with a benchmark of 10 words per minute for *Partially meets minimum proficiency,* 30 words per minute for *Meets minimum proficiency* , and 55 words per minute for *Exceeds minimum proficiency*. In this subdomain, students show a much lower proficiency than for aural comprehension, which is both understandable and congruent with the EGRA results presented above. At the baseline only 4.3% of students met or exceeded minimum proficiency, and the large majority—92.5%—did not meet minimum proficiency. At the end-of-project assessment, 42.5% met or exceeded minimum proficiency (of which 16.8% even exceeded it), while only 28.8% of students did not meet the minimum standard. These changes are statistically significant, and represent a huge increase in average decoding proficiency. Similar improvements could be observed for both boys and girls, again with no statistical difference between genders. The results are presented in Table 17 as well as in Figure 13.

		Does not meet		**Partially meets**		**Meets**		**Exceeds**		**Total**
		%	**95% CI**	**%**	**95% CI**	**%**	**95% CI**	**%**	**95% CI**	
Male	**Baseline**	89.80%	86.44% 92.41%	4.89%	3.02% 7.82%	3.66%	2.32% 5.74%	1.65%	0.81% 3.30%	100%
	End-of-project	27.30%	23.09% 31.94%	29.50%	25.10% 34.22%	24.70%	20.76% 29.10%	18.60%	15.15% 22.51%	100%
	Chi2 Test	F(2.93, 4762.67)= 111.9								
	p-value	0.000								
Female	**Baseline**	95.00%	92.82% 96.56%	1.61%	0.81% 3.18%	1.06%	0.52% 2.16%	2.32%	1.32% 4.05%	100%
	End-of-project	30.40%	26.02% 35.13%	27.90%	23.75% 32.47%	26.60%	22.57% 31.11%	15.10%	12.09% 18.67%	100%
	Chi2 Test	F(2.94, 4852.01)= 158.5								
	p-value	0.000								
All	**Baseline**	92.50%	90.62% 94.08%	3.17%	2.12% 4.73%	2.30%	1.56% 3.38%	2.00%	1.29% 3.09%	100%
	End-of-project	28.80%	25.92% 31.94%	28.70%	25.60% 31.97%	25.70%	22.86% 28.67%	16.80%	14.54% 19.38%	100%
	Chi2 Test	F(2.92, 4826.73)= 257.37								
	p-value	0.000								

Table 17 GPF Results Decoding proficiency, by survey round and gender

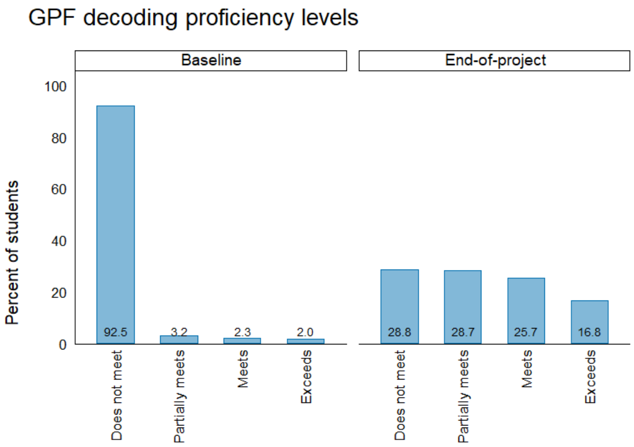
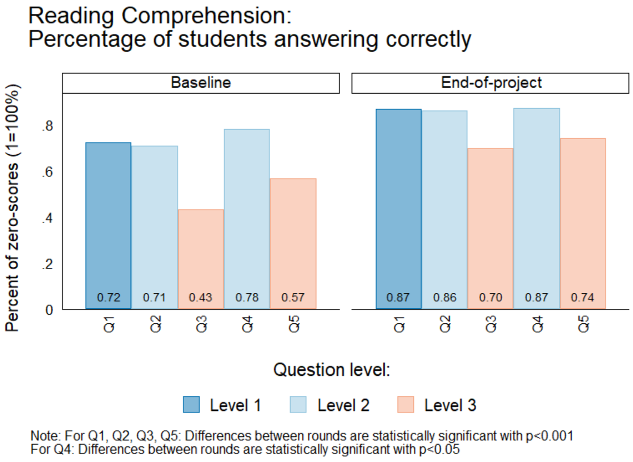


Figure 13 GPF/EGRA Mapping: Decoding results in GPF proficiency levels, by survey round

5.2.3. READING COMPREHENSION OF SIMPLE, GRADE 2-LEVEL CONNECTED TEXT

Finally, the EGRA subtask "reading comprehension" is used to assess students' proficiency levels for the GPF subdomain *"Retrieve information at sentence or text level: Retrieve prominent information when information is found in a single sentence containing no competing information. The information is generally a response to a 'who, what, when and where" question and the information sought is generally names, facts or numbers*". Meets minimum proficiency is here defined as correctly answering to at least three of the five comprehension questions (corresponding to answering correctly all questions corresponding to difficulty/proficiency levels 1 and 2, see annex II). Figure 14 shows the percentage of students answering correctly to any one of the five questions, by the questions' level of difficulty.



Note: For Q1, Q2, Q3, Q5: Differences between rounds are statistically significant with $p<0.001$
For Q4: Differences between rounds are statistically significant with $p<0.05$

Figure 14 Percentage of students answering to correctly to any of the five reading comprehension questions, by level of difficulty

Results are presented in Table 18 as well as in Figure 15. At the baseline, a large majority of learners (84.7%) did not meet minimum proficiency in this subtask, while 5.97% of students met minimum proficiency and 4.8% even exceeded it. By the end of the project, the percentage of students not meeting minimum proficiency had decreased to 23.5%, while 20.6% of learners met minimum proficiency and a full 43.7% even exceeded it. These substantial changes are also statistically significant. Improvements can be observed for both boys and girls, with no statistical differences between the genders.

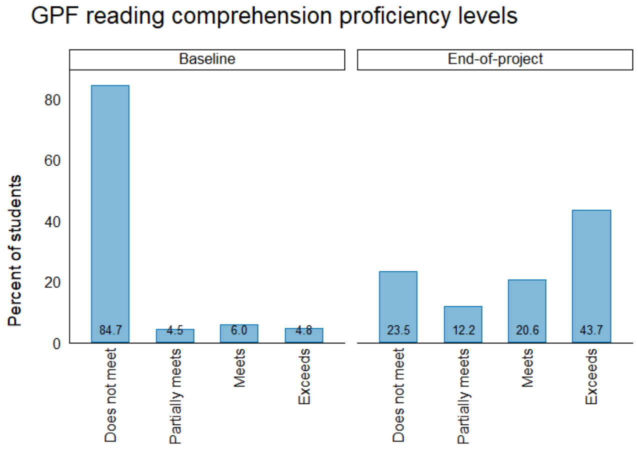


Figure 15 GPF/EGRA Mapping: Reading comprehension results in GPF proficiency levels, by survey round

		Does not meet		Partially meets		Meets		Exceeds		Total
		%	95% CI	%	95% CI	%	95% CI	%	95% CI	
Male	Baseline	82.90%	79.10% 86.06%	5.27%	3.50% 7.86%	5.73%	3.89% 8.37%	6.14%	4.28% 8.75%	100%
	End-of-project	23.80%	19.87% 28.31%	11.50%	8.53% 15.28%	21.10%	17.37% 25.31%	43.60%	39.08% 48.26%	100%
	Chi2 Test	F(2.97, 4878.32)= 100.2								
	p-value	0.000								
Female	Baseline	86.40%	83.13% 89.10%	3.83%	2.46% 5.90%	6.20%	4.36% 8.74%	3.59%	2.28% 5.58%	100%
	End-of-project	23.20%	19.20% 27.67%	12.90%	10.00% 16.36%	20.20%	16.63% 24.39%	43.80%	39.32% 48.28%	100%
	Chi2 Test	F(2.99, 4995.40)= 129.1								
	p-value	0.000								
All	Baseline	84.70%	82.43% 86.74%	4.51%	3.35% 6.06%	5.97%	4.62% 7.69%	4.80%	3.62% 6.35%	100%
	End-of-project	23.50%	21.01% 26.20%	12.20%	10.04% 14.67%	20.60%	18.04% 23.52%	43.70%	40.82% 46.59%	100%
	Chi2 Test	F(2.95, 4961.35)= 237.6								
	p-value	0.000								

Table 18 GPF Results reading Comprehension, by survey round and gender

5.2.4. OVERALL PROFICIENCY LEVELS (ES.1.1)

One of the project's main performance indicators is ES.1.1: *Percent of learners targeted for USG assistance who attain a minimum grade-level proficiency in reading at the end of grade 2*. A student can only be considered to meet the grade level minimum proficiency if he or she performs adequately in all subdomains. Correspondingly, a students' overall proficiency level can be assessed by identifying the minimum proficiency level that he or she meets in each individual subdomain; and whether a student overall meets grade-level proficiency can be determined by whether he or she meets grade-level proficiency in all three domains. Furthermore, it is important to remember in this context that the present evaluation design does not allow to make claims on the project's causal contribution to any learning gains, and that all results presented in this report need to be interpreted as purely descriptive statements on how reading levels changed over the duration of the project (i.e., over the duration of grade 2).

Figure 16, together with Table 19 and Table 20, shows how the intersecting set of students meeting grade-level proficiency across domains has changed between baseline and end-of-project assessment and already indicates that the percentage of students meeting grade-level standards has increased. Table 22 and Figure 18 summarize these results by presenting whether a learner did or did not meet minimum grade level reading proficiency. Table 21, together with Figure 17, shows the results of this overall GPF mapping by proficiency category.

As it turns out, at the baseline a mere 3.33% of learners had met or exceeded minimum grade level reading proficiency, while 94.5% of learners did not meet the standards. At the end of the project, however, 40.9% of learners met or exceeded minimum grade level reading proficiency standards (out of which 13.5% exceeded these standards). Furthermore, even among those not meeting minimum standards at the end of the project, an important group at least partially met the minimum proficiency standards, while only 33.7% did not meet them at all. Again, improvements can be observed for both boys and girls, with no specific gender differences emerging.

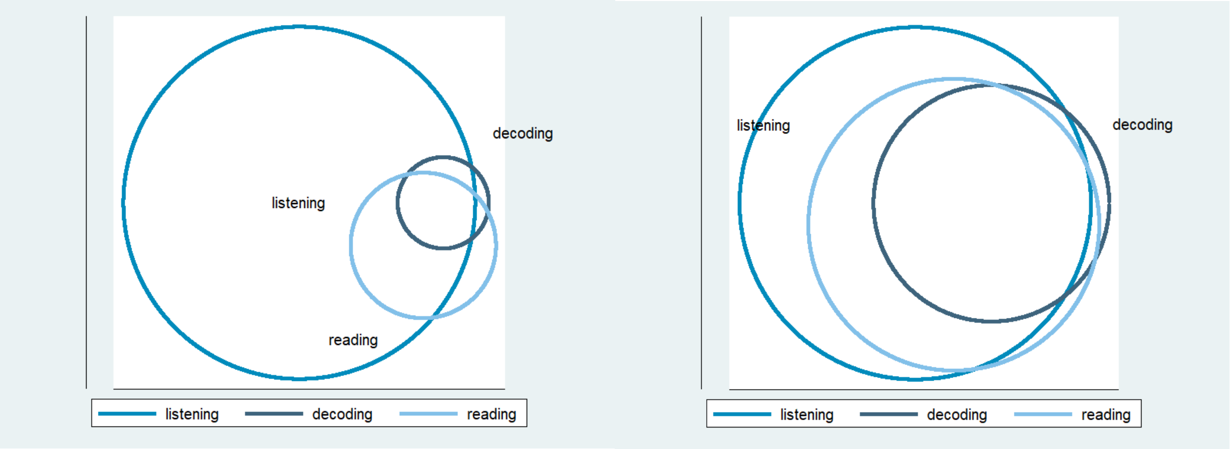


Figure 16 Venn diagram showing the intersecting sets of students meeting grade-level proficiency standards across domains, at the baseline (left) and at the end-of-project (right)

		Listening Comprehension			
		Meets		Does not meet	
		Reading Comprehension		Reading Comprehension	
		Meets	Does not meet	Meets	Does not meet
Decoding	Meets	3.34%	0.80%	0.09%	0.08%
	Does not meet	6.14%	53.73%	1.36%	34.48%

Table 19 Intersecting sets of students meeting grade-level proficiency standards across domains at baseline

		Listening Comprehension			
		Meets		Does not meet	
		Reading Comprehension		Reading Comprehension	
		Meets	Does not meet	Meets	Does not meet
Decoding	Meets	40.90%	0.94%	0.31%	0.34%
	Does not meet	23.28%	29.25%	0.19%	4.80%

Table 20 Intersecting sets of students meeting grade-level proficiency standards across domains at end-of-project assessment

		Does not meet		Partially meets		Meets		Exceeds		Total
		%	95% CI	%	95% CI	%	95% CI	%	95% CI	
Male	Baseline	92.90%	90.15% 94.97%	3.39%	1.98% 5.75%	3.18%	1.96% 5.12%	0.50%	0.16% 1.52%	100%
	End-of-project	32.80%	28.36% 37.64%	25.30%	21.19% 29.99%	27.10%	23.00% 31.51%	14.80%	11.77% 18.41%	100%
	Chi2 Test	F(2.89, 4704.97)= 123.6406								
	p-value	0.000								
Female	Baseline	96.00%	93.94% 97.31%	1.03%	0.44% 2.40%	1.83%	1.03% 3.23%	1.19%	0.52% 2.72%	100%
	End-of-project	34.70%	30.11% 39.50%	25.40%	21.46% 29.73%	27.80%	23.68% 32.28%	12.20%	9.50% 15.53%	100%
	Chi2 Test	F(2.96, 4887.03)= 140.5358								
	p-value	0.000								
All	Baseline	94.50%	92.91% 95.77%	2.15%	1.36% 3.41%	2.47%	1.71% 3.57%	0.86%	0.44% 1.68%	100%
	End-of-project	33.70%	30.81% 36.80%	25.40%	22.47% 28.47%	27.40%	24.58% 30.43%	13.50%	11.45% 15.84%	100%
	Chi2 Test	F(2.94, 4852.54)= 254.3990								
	p-value	0.000								

Table 21 GPF Results: Overall proficiency levels, by survey round and gender

	Male		Female		All	
	%	95% CI	%	95% CI	%	95% CI
Baseline	3.68%	2.36% 5.67%	3.02%	1.89% 4.81%	3.34%	2.42% 4.57%
End-of-project	41.80%	37.27% 46.53%	40.00%	35.45% 44.67%	40.90%	37.91% 43.96%
Difference, absolute	38.12%***		36.98%***		37.56%**	
Difference, relative	1036%		1225%		1125%	
Chi2 Test	F(1, 1626) = 216.90		F(1, 1649) = 222.39		F(1, 1651) = 464.57	
p-value	0.000		0.000		0.000	

Table 22 GPF Results: Percentage of learners meeting overall grade-level proficiency, by survey round and gender

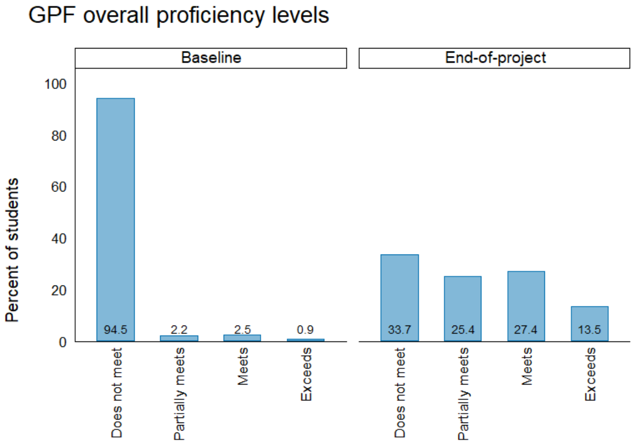


Figure 17 GPF/EGRA Mapping: Overall results in GPF proficiency levels, by survey round

Finally, as this proficiency level assessment is the overall proficiency indicator used by the SENSE program, Figure 19 to Figure 22 present the results by gender and state, both in proficiency categories and in the simplified met/did-not-meet-standards format. As has become abundantly clear, no gender differences emerge. However, the data shows that most of the improvements were concentrated in Gombe State: Here, the percentage of those meeting minimum proficiency increased from 0.4% at the baseline to 64.6% at the end-of-project assessment. Meanwhile, in Adamawa, the percentage of those meeting minimum proficiency increased from 6.6% at the baseline to 16.9% at the end of the project. As is also shown in Table 23, even this relatively speaking poorer performance of learners in Adamawa presents an impressing change of 156% in the percentage of those meeting grade-level proficiency.

	Adamawa		Gombe		All	
	%	95% CI	%	95% CI	%	95% CI
Baseline	6.59%	90.93% 95.25%	0.41%	98.33% 99.90%	3.34%	2.42% 4.57%
End-of-project	16.90%	79.07% 86.42%	64.60%	31.20% 39.86%	40.90%	37.91% 43.96%
Difference, absolute	10.31%***		64.19%***		37.56 %**	
Difference, relative	156%		15656%		1125%	
Chi2 Test	F(1, 776) =	24.936	F(1, 875) =	431.81	F(1, 1651) =	464.57
p-value	0.000		0.000		0.000	

Table 23 GPF Results: Percentage of learners meeting overall grade-level proficiency, by survey round and state

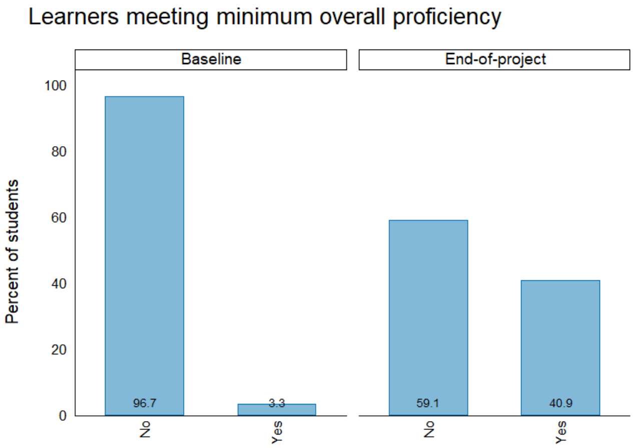


Figure 18 GPF/EGRA Mapping: Percentage of learners meeting minimum grade-level reading proficiency, by survey round

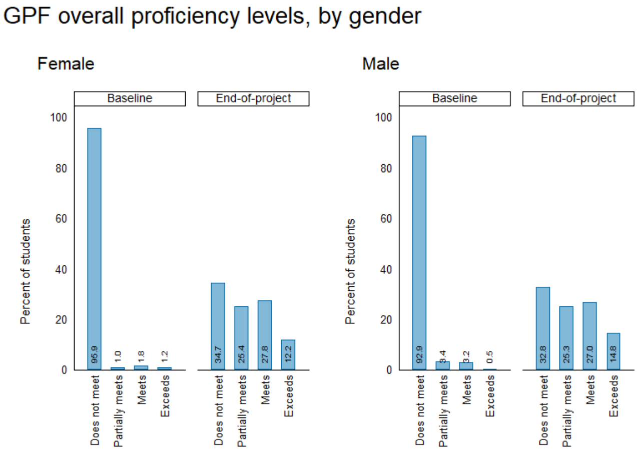


Figure 19 GPF/EGRA Mapping: Overall results in GPF proficiency levels, by survey round and gender

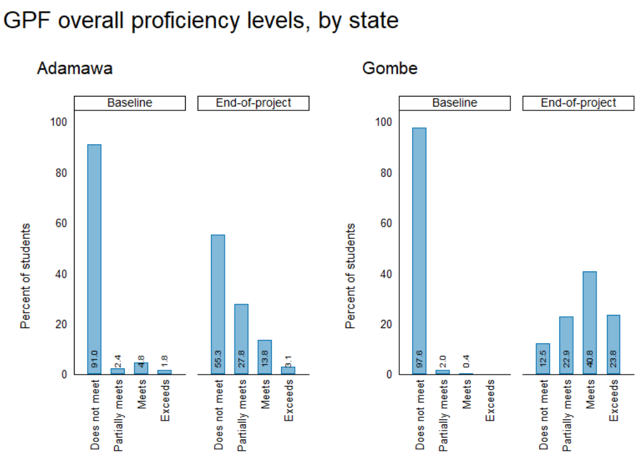


Figure 20 GPF/EGRA Mapping: Overall results in GPF proficiency levels, by survey round and state

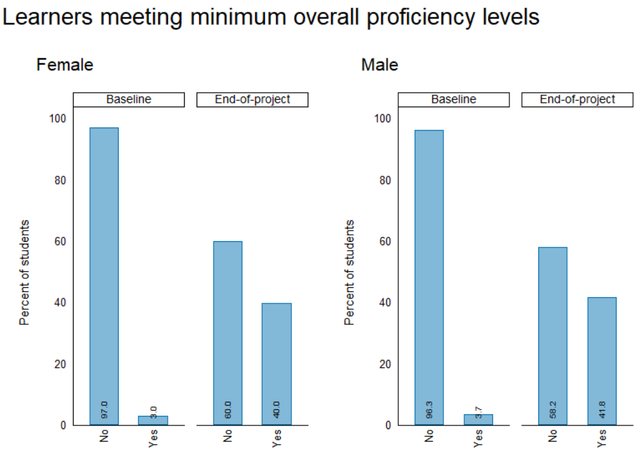


Figure 21 GPF/EGRA Mapping: Percentage of learners meeting minimum grade-level reading proficiency, by survey round and gender

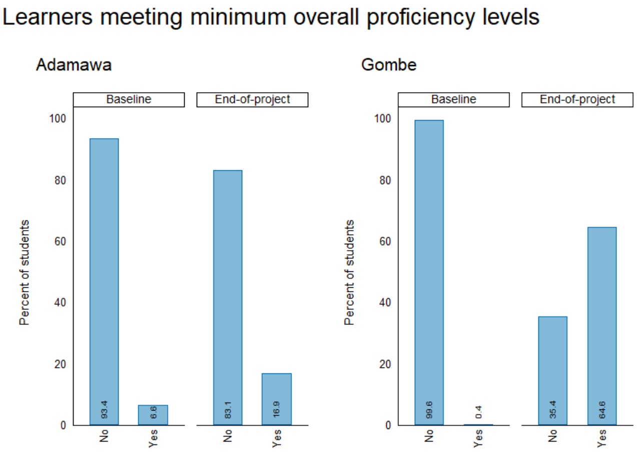


Figure 22 GPF/EGRA Mapping: Percentage of learners meeting minimum grade-level reading proficiency, by survey round and state

5.2.5. CHANGES IN OVERALL PROFICIENCY LEVELS (ES.1-48)

A further important project performance indicator is ES.1-48: *Percent of learners targeted for USG assistance with an increase of at least one proficiency level in reading at the end of grade 2*. As described in sections 2.2 and 3, the project evaluation strategy had to be changed, and it was not possible for the project team to re-assess the same set of students that had been assessed at baseline. Without such panel data, it is not possible to determine exactly which of the students falling in a given proficiency level at the project end had improved and by how much, which students had stayed at the same level, and which students have deteriorated. However, it is possible to determine based on the data the minimum number of students who must have, and the maximum number of students who could have, improved in order to see the observed change in the distribution across proficiency levels.

The results of this analysis are presented in Table 24. In addition to the logically possible minimum and maximum number of students who improved the table also contains the arithmetic average between these two numbers as the best guess on the actual number of students who improved. The results show that around 63.5% of students managed to improve by at least one proficiency level in reading over the project duration. The error range of that estimate is large (from around 51.5% to around 72%), yet even the lower end of that estimate shows substantial learning gains.

	Minimum	95% CI	Maximum	95% CI	Arithmetic midpoint
Male	60.63%	45.09% 75.97%	66.67%	56.45% 77.48%	63.65%
Female	62.54%	49.01% 76.07%	64.16%	55.91% 72.83%	63.35%
All	61.68%	51.52% 71.67%	65.42%	59.18% 71.94%	63.55%

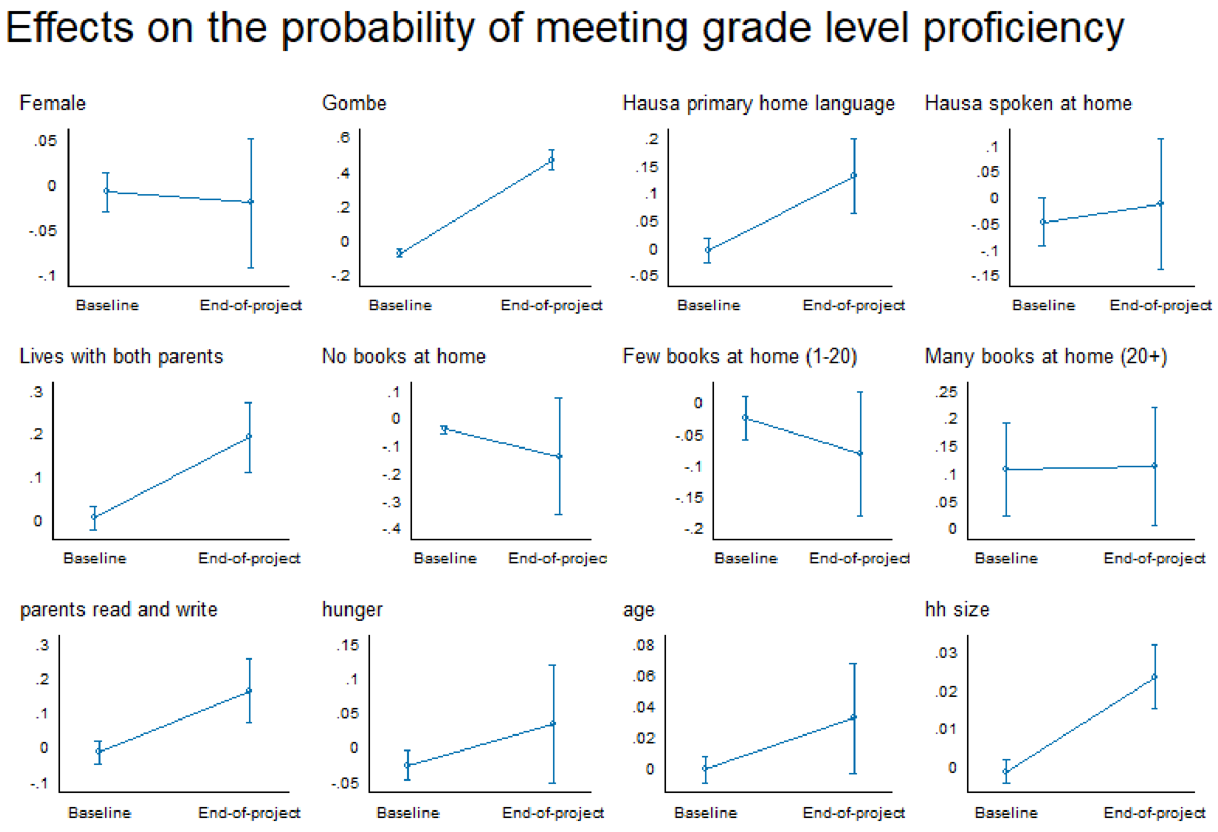
Table 24 Estimated percentage of students who improved by at least one proficiency level in reading at the end of grade 2

6. STUDENT CHARACTERISTICS INFLUENCING IMPROVEMENT RATES

This section takes a closer look at the project's main outcomes of interest (ES.1.1: *Percent of learners targeted for USG assistance who attain a minimum grade-level proficiency in reading at the end of grade 2*, and ES.1-48: *Percent of learners targeted for USG assistance with an increase of at least one proficiency level in reading at the end of grade 2*), analyzing whether learners with specific characteristics were more likely to show learning gains. Again, the depth of the analysis is limited by the lack of a comparison group and panel data, but some general statements can be made. The data is analyzed by first calculating correlations between independent variables of interest and baseline and end-of-project surveys, respectively. In a second step, logit estimations are used to estimate the effect of the independent variables of interest on the probability of meeting grade-level standards. The regressions take the following form:

$$P(meeting_{GP}) = \beta_0 + \beta_1 round + \beta_2 indep\_var^i + \beta_3 round * indep\_var^i$$

Results are reported in Table 25 and graphically (together with error margins) in Figure 23. The results uncover few characteristics that determine learning outcomes. The only factor that strongly matter more at the end-of-project than at the baseline is living in Gombe. Additionally, there is some evidence (at the 5% significance level) that living in a home where Hausa is spoken increases learning gains, and that learning gains might be larger for children from more disadvantaged backgrounds (those having no books at home, who went hungry in the past week, or living in a larger household). Having parents who can read and write also seems to be associated with larger learning gains (at the 5% significance level). Also notably, and in line with the previously reported results, gender does not affect the probability of meeting grade-level proficiency.



Note: Not living with parents completely determines failure to meet standards at baseline

Figure 23 Effects on the probability of meeting grade level proficiency

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Indep. Var.	**Gender = female**	**State = Gombe**	**Hausa primary home language**	**Hausa spoken at home**	**Lives with any parent**	**Lives with both parents**	**No books at home**	**Few books at home (1-20)**	**Many books at home (21+)**	**Parents read and write**	**Went hungry in past week**	**Age**	**House-hold size**
(I)	GPF	-0.05	-0.477***	-0.013	-0.145	0.044***	0.028	-0.113***	-0.106	0.219**	-0.038	-0.146*	0.024	-0.513
		0.560	0.000	0.870	0.060	0.000	0.620	0.000	0.180	0.010	0.500	0.010	0.910	0.460
	Const.	0.525***	0.543***	0.609***	0.885***	0.956***	0.842***	0.136***	0.794***	0.070***	0.876***	0.272***	7.813***	8.265***
		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(II)	GPF	-0.019	0.493***	0.135***	-0.004	0.057***	0.114***	-0.012	-0.044	0.056*	0.084***	0.024	0.125	1.676***
		0.600	0.000	0.000	0.870	0.000	0.000	0.200	0.100	0.030	0.000	0.400	0.070	0.000
	Const.	0.507***	0.301***	0.521***	0.906***	0.914***	0.781***	0.027***	0.859***	0.114***	0.825***	0.199***	8.747***	8.211***
		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(III)	a. Project End	2.936***	1.062***	2.633***	2.219***	3.007***	2.444***	2.785***	2.751***	3.099***	2.073***	2.762***	1.884	1.825***
		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.130	0.000
	b. Indep. Var.	-0.202	-2.829***	-0.055	-0.996*	1.149**	0.226	-1.920***	-0.559	1.690***	-0.315	-0.952	0.015	-0.031
		0.560	0.000	0.870	0.010	0.000	0.650	0.000	0.140	0.000	0.460	0.070	0.910	0.520
	c. Inter-action	0.125	5.020***	0.618	0.954*		0.652	1.320*	0.235	-1.224**	1.061*	1.100*	0.125	0.135*
		0.740	0.000	0.100	0.050		0.230	0.020	0.590	0.010	0.030	0.050	0.410	0.010
	Const.	-3.266***	-2.652***	-3.334***	-2.549***	-4.469***	-3.558***	-3.140***	-2.847***	-3.532***	-3.086***	-3.175***	-3.485***	-3.118***
		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: (I) and (II) are correlations between “Meets grade level proficiency” (GPF) and the respective independent variable at the baseline and end-of-project assessment, respectively. Numbers reported are estimation coefficients and p-values. (III) are the odds-ratios derived from logit-estimation models that regress the survey round, the respective independent variable, and an interaction between the two on whether the student met grade level proficiency. Reported numbers are odds-ratios and p-values. Odds-ratios above 1 indicate a positive influence, odds-ratios below 1 indicate a negative influence. All estimates take into account school clusters and sampling weights. Living with any or both parents fully predicts not meeting grade level proficiency at the baseline, hence no interaction can be calculated.

Interpretation example for (2): (I) At the baseline, living in Gombe is negatively correlated with meeting grade level standards. (II) However, and the end-of-project assessment, living in Gombe is positively correlated with meeting grade-level standards. (III) (a) Controlling for living in Gombe, the odds of meeting grade level standards are slightly higher at the end-of-project compared to the baseline; (c) for students living in Gombe, the odds at the end-of-project are 5 times as high compared to students from Adamawa at the baseline . (b) At the same time, at the baseline the odds are three times as high for meeting grade level standards for students living in Adamawa compared to those living in Gombe.

Significance levels: *** p≤0.001 ; ** p≤0.01 ; * p≤0.05

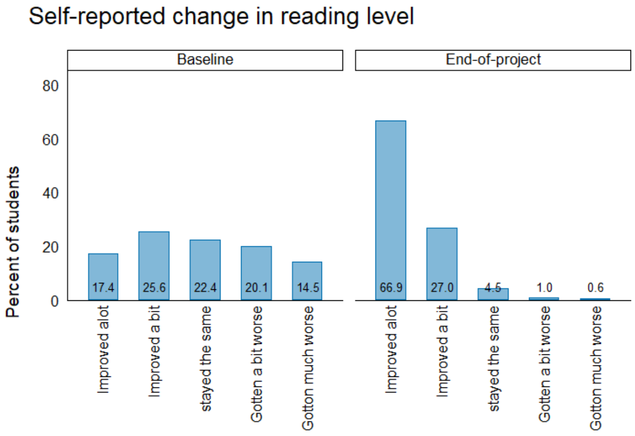
Table 25 Factors influencing meeting minimum grade-level proficiency

7. STUDENT SELF-REPORTS

To complement the results of the EGRA assessment, triangulate the results, and obtain some further insights into the perceived benefits of the program, participants were asked a set of complementary questions after completing the EGRA assessment. Questions concerned the perceived learning gains in the six months prior to the assessment; whether the child enjoys learning to read; and the perceived usefulness of the available learning materials. Questions were asked at the baseline and at the end-of-project survey in order to allow for a comparison. More so than for the results of the EGRA assessment, it is likely that any changes in the answers are due to program exposure, even though in the absence of a comparison group no causal claims can be made. The likely attributability of the results to the SENSE program arises from the supposition that the general learning environment, apart from the project exposure, did not change, and that any potential biases in students' answers were similarly likely to appear at the baseline assessment as well and can thus be "netted out". The results are presented in the following and provide supporting evidence for the effectiveness of the SENSE approach.

7.1. CHANGE IN PERCEIVED READING PROFICIENCY

Participants' perceptions on how their reading level changed over the six month before the assessment are in line with the EGRA results presented above: As shown in Table 26 and Figure 24, at the end of the project, two thirds (66.9%) of students thought their reading improved a lot, and more than a quarter (27%) thought they improved a bit. Less than 2% thought it got worse, and 4.5% thought it stayed the same. By comparison, at the baseline only an eighth of students (17.4%) thought they had improved a lot, and a quarter (25.6%) thought they improved a bit; more than a third of students (35.5%) thought they had even gotten worse at reading. These differences between the survey rounds are highly statistically significant. Notably, a disaggregation of the results by gender and state mirrors the EGRA results: Boys' and girls' perceptions about their learning gains evolved in a similar fashion, with no differences between the genders being detectable statistically. However, in line with the diverging EGRA improvement rates across states, students' perceived learning gains are much larger in Gombe than in Adamawa (in Gombe, 80.8% of students thought they improved a lot, while the same is true for "only" 51.9% in Adamawa). This despite the fact that at the baseline, students in Adamawa had thought that they had improved more – which is also in line with the better EGRA baseline results in that state. It is important to point out, however, that in both states, the vast majority of students thought they improved at least a bit: 97.4% in Gombe, and 90.4% in Adamawa.



Prompt: Over the past 6 months, do you think your reading level has…

Figure 24 Self-reported change in the reading level, by survey round

		Improved a lot		Improved a bit		Stayed the same		Got a bit worse		Got much worse		Total
		%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI	
All	Baseline	17.4%	15.1% 20.0%	25.6%	22.7% 28.7%	22.5%	19.7% 25.5%	20.1%	17.5% 23.0%	14.4%	12.2% 17.1%	100%
	End-of-project	66.9%	63.7% 70.0%	27.0%	24.1% 30.0%	4.5%	3.2% 6.3%	1.0%	0.5% 2.2%	0.6%	0.2% 1.8%	100%
	Chi2 Test	F(3.83, 5773.10)= 120.2										
	p-value	0.000										
Male	Baseline	19.8%	16.2% 24.0%	25.6%	21.4% 30.3%	19.1%	15.5% 23.3%	21.8%	18.1% 26.2%	13.7%	10.4% 17.8%	100%
	End-of-project	66.5%	61.7% 70.9%	27.0%	22.8% 31.6%	4.3%	2.7% 6.9%	1.5%	0.6% 3.6%	0.7%	0.2% 3.2%	100%
	Chi2 Test	F(3.77, 5855.40)= 51.0										
	p-value	0.000										
Female	Baseline	15.2%	12.2% 18.8%	25.7%	21.7% 30.0%	25.5%	21.6% 30.0%	18.4%	15.0% 22.5%	15.2%	12.1% 18.9%	100%
	End-of-project	67.4%	62.7% 71.7%	27.0%	23.0% 31.4%	4.7%	2.9% 7.6%	0.6%	0.2% 2.0%	0.4%	0.1% 1.9%	100%
	Chi2 Test	F(3.95, 6178.34)= 78.2										
	p-value	0.000										
Adamawa	Baseline	26.4%	22.5% 30.7%	23.7%	19.9% 28.1%	13.9%	10.7% 17.8%	18.9%	15.4% 23.0%	17.1%	13.7% 21.3%	100%
	End-of-project	51.9%	46.7% 57.0%	38.5%	33.6% 43.6%	7.1%	4.7% 10.6%	1.7%	0.7% 4.0%	0.9%	0.2% 3.8%	100%
	Chi2 Test	F(3.67, 2498.47)= 31.2										
	p-value	0.000										
Gombe	Baseline	9.5%	7.0% 12.8%	27.3%	23.2% 31.7%	30.0%	25.8% 34.5%	21.1%	17.5% 25.3%	12.1%	9.3% 15.7%	100%
	End-of-project	80.8%	77.1% 84.0%	16.4%	13.3% 19.9%	2.1%	1.2% 3.9%	0.4%	0.1% 1.6%	0.3%	0.1% 1.0%	100%
	Chi2 Test	F(3.79, 3136.98)= 147.5										
	p-value	0.000										

Table 26 Self-reported changes in the reading level over the past six months, by survey round, gender, and state

7.2. ENJOYMENT OF READING

Learning to read should not be a chore; it should be something that children enjoy, as it supports children in their natural desire to explore and understand the world around them. It therefore is important that children perceive reading as something fun. Participants were thus asked, both at the baseline and at the end-of-project assessment, whether they enjoy learning to read and want to get better at it. Already at the baseline, 63.8% of learners agreed completely, and a further 20.1% agreed a little bit. 8.1% did not like it much, and 2.7% not at all. These numbers improved by the end-of-project survey, when 72.9% of learners completely agreed, and 23.2% agreed a little bit. Only 2.4% agreed not much, and 0.2% agreed not at all (see Table 27 and Figure 25). These changes are statistically significant and indicate that the playful approach of the SENSE program might help to improve children's attitude towards reading (though such causality cannot be established based on the available data).

Again, similar patterns emerge for boys and girls, with both genders' attitudes evolving in a similar way, and no statistical differences emerging between genders. However, similar as for self-reported changes in the reading levels, differences do emerge between Adamawa and Gombe. In Gombe, the percentage of students reporting to enjoy learning to read increased from 54.9% at the baseline to 83.6% and the end-of-project survey; the percentage of those reporting to at least enjoy it a little bit increased from 75.5% to 98.3%. At the same time, the percentage of students saying they completely enjoy learning to read actually decreased, from 74.1% to 61.2%, and the percentage of those reporting to at least enjoy it a little did not change (the change from 93.7% to 93.5% is not statistically significant).

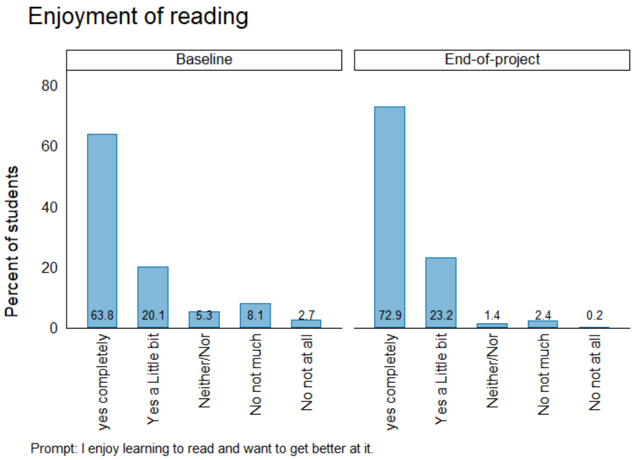


Figure 25 Enjoyment of reading, by survey round

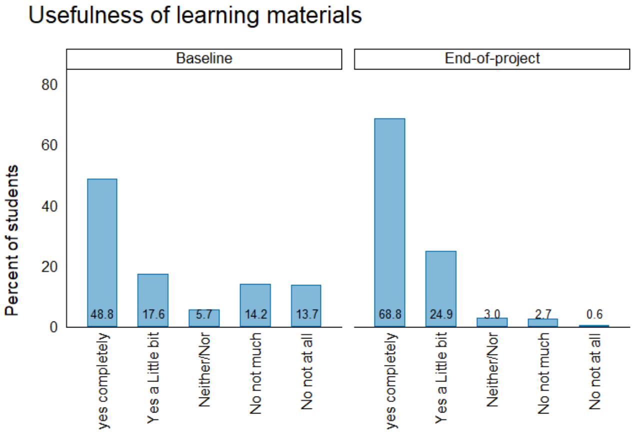
		Yes, completely		Yes, a little bit		Neither/Nor		No, not much		No, not at all		Total
		%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI	
All	**Baseline**	63.8%	60.6% 67.0%	20.1%	17.5% 23.1%	5.3%	3.9% 7.2%	8.1%	6.4% 10.1%	2.7%	1.8% 4.1%	100%
	End-of-project	72.9%	69.9% 75.6%	23.2%	20.4% 26.1%	1.4%	0.8% 2.6%	2.4%	1.6% 3.5%	0.2%	0.0% 1.4%	100%
	Chi2 Test	F(3.71, 5627.57)=	15.10									
	p-value	0.000										
Male	**Baseline**	64.2%	59.3% 68.8%	19.7%	15.9% 24.0%	4.3%	2.5% 7.1%	8.4%	6.1% 11.6%	3.5%	2.0% 5.8%	100%
	End-of-project	72.5%	67.8% 76.7%	23.2%	19.2% 27.8%	1.9%	0.9% 3.8%	2.0%	1.0% 4.2%	0.4%	0.1% 2.7%	100%
	Chi2 Test	F(3.88, 6048.54)=	6.86									
	p-value	0.000										
Female	**Baseline**	63.5%	58.8% 67.9%	20.6%	16.9% 24.9%	6.3%	4.3% 9.1%	7.7%	5.5% 10.8%	2.0%	1.0% 3.8%	100%
	End-of-project	73.2%	68.8% 77.3%	23.1%	19.3% 27.4%	1.0%	0.3% 2.7%	2.7%	1.6% 4.6%	0.0%		100%
	Chi2 Test	F(3.96, 6215.73)=	9.51									
	p-value	0.000										
Adamawa	**Baseline**	74.1%	69.6% 78.2%	19.6%	15.9% 23.9%	1.9%	1.0% 3.7%	2.1%	1.1% 3.9%	2.3%	1.2% 4.3%	100%
	End-of-project	61.2%	56.4% 65.8%	32.3%	27.8% 37.2%	1.4%	0.6% 3.5%	4.7%	3.1% 7.0%	0.4%	0.1% 3.0%	100%
	Chi2 Test	F(3.72, 2537.28)=	5.85									
	p-value	0.000										
Gombe	**Baseline**	54.9%	50.4% 59.4%	20.6%	17.0% 24.8%	8.2%	5.8% 11.4%	13.2%	10.4% 16.8%	3.0%	1.8% 5.2%	100%
	End-of-project	83.6%	80.0% 86.6%	14.7%	11.8% 18.2%	1.4%	0.7% 3.0%	0.3%	0.0% 1.4%	0.0%		100%
	Chi2 Test	F(3.96, 3313.34)=	31.26									
	p-value	0.000										

Table 27 Reported enjoyment of reading, by survey round, gender, and state

7.3. USEFULNESS OF LEARNING MATERIALS

The third topic that learners were asked concerns whether they find the available learning materials useful to learn and practice to read. At the end of the project, learners were much more likely completely or at least somewhat agree (see Table 28 and Figure 26): 48.8% completely agreed at the baseline, and 66.4% at least agreed a little bit; at the end-of-project assessment, the respective percentages were 68.8% and 93.7%. These changes are substantial and statistically significant. They are an indication that the learning materials developed for the SENSE program might not just be effective in improving learners' enjoyment of reading (see last section), but are also perceived as helpful by children, which then might translate into improved reading proficiency (see the reported EGRA results).

As was the case for self-reported learning gains and enjoyment of reading, there are no differences between boys and girls when it comes to the perceived usefulness of learning materials. However, differences do again emerge between Gombe and Adamawa, in a pattern that by now is familiar: At the baseline, students in Adamawa were more likely to completely agree (59%) or at least somewhat agree (75.9%) that the materials available to them were useful to learn to read than in Gombe, where the respective numbers were 39.7% and 57.9%. By the end of the project, however, the situation had changed: in Adamawa, the percentage of those completely agreeing remained unchanged, and the percentage of those at least somewhat agreeing improved to 90.3%. At the same time, numbers improved even more in Gombe, where now 77.1% completely agreed, and 96.8% at least somewhat agreed that the learning materials available to them are useful.



Prompt: The learning materials I have are useful to learn and practice to read.

Figure 26 Perceived usefulness of learning materials, by survey round

		Yes, completely		Yes, a little bit		Neither/Nor		No, not much		No, not at all		Total
		%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI	
All	Baseline	48.8%	45.5% 52.1%	17.6%	15.0% 20.5%	5.7%	4.2% 7.6%	14.2%	12.0% 16.9%	13.7%	11.5% 16.3%	100%
	End-of-project	68.8%	65.6% 71.8%	24.9%	22.1% 27.9%	3.1%	2.0% 4.6%	2.7%	1.7% 4.1%	0.6%	0.3% 1.6%	100%
	Chi2 Test	F(3.96, 5852.55)= 49.0										
	p-value	0.000										
Male	Baseline	47.7%	42.8% 52.6%	17.9%	14.2% 22.2%	4.9%	3.2% 7.7%	15.0%	11.8% 19.0%	14.5%	11.2% 18.6%	100%
	End-of-project	68.2%	63.5% 72.5%	25.1%	21.1% 29.6%	3.1%	1.8% 5.5%	2.7%	1.4% 4.9%	0.9%	0.3% 2.7%	100%
	Chi2 Test	F(3.93, 6075.13)= 24.2										
	p-value	0.000										
Female	Baseline	49.8%	45.2% 54.5%	17.3%	13.9% 21.3%	6.3%	4.3% 9.2%	13.5%	10.5% 17.2%	13.0%	10.1% 16.7%	100%
	End-of-project	69.4%	64.6% 73.8%	24.6%	20.5% 29.2%	3.0%	1.6% 5.3%	2.6%	1.5% 4.6%	0.4%	0.1% 1.4%	100%
	Chi2 Test	F(3.85, 5959.37)= 27.8										
	p-value	0.000										
Adamawa	Baseline	59.0%	54.2% 63.7%	16.9%	13.3% 21.1%	2.7%	1.5% 4.6%	7.3%	5.0% 10.5%	14.2%	11.0% 18.1%	100%
	End-of-project	59.7%	54.8% 64.4%	30.6%	26.1% 35.4%	3.9%	2.2% 6.9%	4.5%	2.9% 7.0%	1.4%	0.6% 3.3%	100%
	Chi2 Test	F(3.96, 2637.08)= 14.7										
	p-value	0.000										
Gombe	Baseline	39.7%	35.4% 44.1%	18.2%	14.7% 22.3%	8.3%	6.0% 11.5%	20.5%	16.9% 24.5%	13.3%	10.4% 16.9%	100%
	End-of-project	77.1%	73.1% 80.7%	19.7%	16.3% 23.6%	2.3%	1.2% 4.1%	1.0%	0.3% 3.0%	0.0%	0.0%	100%
	Chi2 Test	F(3.84, 3121.13)= 47.3										
	p-value	0.000										

Table 28 Perceived usefulness of available reading learning materials

8. DISCUSSION OF RESULTS

The analysis presented in this report shows impressive learning advances over the duration of the project. This is true for all EGRA subtasks, and, by extension, for the measure of grade-level proficiency that was derived from the EGRA data. The scale of the learning advances is impressive, raising the question of whether we can trust the results, and if so, how such changes may be explained.

8.1. CAN WE TRUST THE RESULTS?

There are two aspects to the question of whether we can trust the results: First, is the data itself reliable? And second, to what extent do the results reflect the effect of the SENSE program, as opposed to other causal influences?

First, as outlined in the methodology section (section 3), enumerators were thoroughly trained in the application of the assessment, data quality checks were in place during the data collection process in the form of close field supervision and an instant or quasi-instant review of the collected data, and the end-of-project data collection was carried out by an independent consultant to minimize conflicts of interests. The assessment data should therefore be of good quality. Triangulation consideration support this notion, as the data and results are consistent across survey modules and data sets: On the one hand, the EGRA results are mirrored by students' self-reported learning gains, including as far as differences across States and a lack of differences across genders is concerned. On the other hand, the stark differences between Adamawa and Gombe State are consistent with a—separate—assessment of School-Based Management Committees (SBMC), which documented generally engaged leadership and strong SBMCs in Gombe and a lack thereof in Adamawa.

Second, the question of the plausibility of a causal contribution of the SENSE activity is much harder to answer. As pointed out in section 2.2, due to various factors no data could be collected in non-participating schools, making it impossible to statistically separate learning gains that are due to the SENSE project from learning gains due to other factors, such as simply progressing through the standard grade curriculum. As was pointed out before, some improvements in reading skills are to be expected even in the absence of the SENSE program, as students at the end of grade 2 will usually have better reading skills than students and the beginning of grade 2. Hence, while it surely would be wrong to claim that all documented learning gains can be attributed to the SENSE project, a comparison of the end-of-project assessment results with the results of other reading assessments in Nigeria shows that students' proficiency at the end of the SENSE activity is markedly above the measured proficiency in other assessments (RTI International 2014; USAID n.d.; n.d.). Furthermore, the strong difference between Gombe and Adamawa might again be an indication for a causal contribution of the SENSE activity – especially so as at the baseline, students in Gombe actually performed worse than students in Adamawa. There is no reason to believe that even in the absence of the SENSE activity, grade-2 students in Gombe always would show much larger learning gains than students in Adamawa. In conclusion, while the evaluation approach does not allow to quantify the causal contribution of the SENSE activity to the learning gains, a range of qualitative considerations support the idea that the SENSE activity was in fact responsible for a part of the learning gains.

8.2. POTENTIAL REASONS FOR LEARNING GAINS IN THE SENSE PROJECT

The SENSE Activity was significantly different from similar projects in terms of the inputs it provides to the education system to improve reading outcomes for learners. In addition to the 'usual' inputs, such as teachers, provision of teaching and learning materials as well as engagement of parents, the SENSE activity developed other innovative approaches that the SENSE team believes were majorly responsible for the EGRA achievement. These innovations are described below and together they provided the learners the critical push to improve reading.

1. **Teacher resource centers (TRC) availability for teacher training and creating teaching and learning aids for reading.** The establishment of teacher resource centers gave teachers a space where they visit and find resources to create interactive TLMs. The flashcards and posters created by teachers at the TRCs make learning fun, interactive, and meaningful for learning. In addition, SENSE Activity has organized termly boot camps at the teacher resource centers. During these boot camps, teachers are guided to make all the manipulatives they will need for the term. The workshop is also used to provide training that will correct the gaps noticed during a classroom observation.
2. **Supplementary readers to practice reading**. In addition to TLMs provided to all learners in target schools, SENSE provided a set of Hausa language supplementary readers in the hands of learners allowing them to practice reading in addition to the Mu Karanta and or RANA workbooks they used in class.
3. **Intensive, site based coaching and mentoring to support teachers to apply teaching methods that promote reading.** Working with trained SSOs, SENSE was able to be on the ground in classrooms to support teachers. The SSOs supported by SENSE conduct beginning of term meetings with all SENSE focus schools head teachers to discuss how they should support teachers to engage learners. This additional support to teachers from head teachers ensured that teachers actually practiced what they were taught and learnt during TPD session.
4. **Transactional Radio Instruction during the period of lock down due to COVID-19**. This radio-base program greatly reduced loss of instruction. EGRA assessment data (from a separate assessment, not the subject of this present report) indicates that learners therefore did not lose so much instruction time when they were in primary 1.
5. **SBMC activities provided opportunities to train and sensitize parents to support their learners to read at home**. Learners had their own supplementary readers that they took home and read to their parents who encouraged them to read more. Learning therefore did not stop at school only, but continued at home.
6. **Reduction in teacher and learner absence from school through the monitoring conducted by project-initiated community education volunteers**. These volunteers are members of SBMCs trained to monitor and report on teacher and learner absences. This greatly reduced loss of instruction time that would have occurred when teachers are absent from school and class.

9. REFERENCES

AUN. 2019. “Learners Performace Assessment Report: Primary 2 & 4 in Adamawa and Gombe States.” Yola, Nigeria: American University of Nigeria.

———. 2020a. “Sense Project Baseline Reading Proficiency Assessment: Results of the Hausa Early Grade Reading Assessment (EGRA) in Adamawa and Gombe.” Yola, Nigeria: American University of Nigeria.

———. 2020b. “Sense Project Second Baseline Reading Proficiency Assessment: Results of the Hausa Early Grade Reading Assessment (EGRA) in Adamawa and Gombe.” Yola, Nigeria: American University of Nigeria.

———. 2021. “Sense Project End-of-Project Survey, Reading Proficiency Assessment: Results of the Hausa Early Grade Reading Assessment (EGRA) in Adamawa and Gombe.” Yola, Nigeria: American University of Nigeria.

Creative Associates International. 2018. “Northern Education Initiative Plus: Early Grade Reading Assessment Midline Report.” https://pdf.usaid.gov/pdf_docs/PA00WGT7.pdf.

Dynamic Measurement Group Inc. 2010. “DIBELS Next Benchmark Goals and Composite Score.” https://dibels.uoregon.edu/docs/DIBELSNextFormerBenchmarkGoals.pdf.

Malawi Ministry of Education, Science and Technology. 2014. “Proposing Benchmarks for Early Grade Reading in Malawi.” Lilongwe: Malawi Ministry of Education, Science and Technology.

Management Systems International (MSI). 2020. “DRAFT Nigeria Policy Linking Pilot Workshop Report: Setting Global Benchmarks for Grades 2 and 3 Early Grade Reading Assessments in Hausa Language.” Abuja, Nigeria.

RTI International. 2009. “Early Grade Reading Assessment (EGRA) Toolkit.” Research Triangle Park, NC: RTI International.

———. 2014. “Nigeria Reading and Access Research Activity: Results of the 2014 Hausa and English Early Grade Reading Assessments (EGRAs) in Government Primary Schools and IQTE Centers of Jigawa, Kaduna, Kano, and Katsina States.” Research Triangle Park, NC: RTI International.

———. 2016. “Early Grade Reading Assessment (EGRA) Toolkit, Second Edition.” Washington, D.C.: United States Agency for International Development.

USAID. 2011. “Early Grade Reading Assessment: Student Response Form Administrator Instructions and Protocol, 2011.”

———. n.d. “Early Grade Reading Barometer: Nigeria - Bauchi Benchmarks, 2013.” Accessed March 10, 2020a. https://earlygradereadingbarometer.org/nigeria-bauchi/benchmarks.

———. n.d. “Early Grade Reading Barometer: Nigeria - Sokoto Benchmarks, 2013.” Accessed March 10, 2020b. https://earlygradereadingbarometer.org/nigeria-sokoto/benchmarks.

USAID, UNESCO, UIS, DFID, Gates Foundation, ACER, and Catholic University of Uruguay. 2019. “Global Proficiency Framework for Reading and Mathematics, Grades 2 to 6.” United States Agency for International Development (USAID) in collaboration with the United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute of Statistics (UIS); the United Kingdom’s Department for International Development (DFID); the Gates Foundation; the Australian Council for Educational Research (ACER); the Catholic University of Uruguay.

10. ANNEX I: SAMPLE SIZE DISTRIBUTION ACROSS STATES

Adamawa State

S/N	LGA	# of Intervention schools	Sample schools per LGA using 95% CI	Sample of learners per LGA
1	Fufore	22	13	46
2	Ganye	24	15	67
3	Guyuk	19	11	26
4	Hong	24	15	34
5	Michika	24	13	38
6	Mubi South	20	11	51
7	Numan	24	15	32
8	Shelleng	24	15	43
9	Song	24	15	37
10	Toungo	17	10	19
11	Yola North	24	15	88
	TOTAL	246	150	**481**

Gombe State

S/N	LGA	# of Intervention schools	Sample schools per LGA using 95% CI	Sample of learners per LGA
1	Akko	12	9	45
2	Balanga	7	6	8
3	Billiri	10	8	43
4	Dukku	7	6	18
5	Funakaye	7	6	62
6	Gombe	12	9	224
7	Kaltungo	7	6	12
8	Kwami	7	6	20
9	Nafada	7	6	16
10	Shongom	6	4	5
11	Yamaltu Deba	7	6	30
	TOTAL	89	72	**483**

11. ANNEX II ORAL READING AND LISTENING COMPREHENSION: STUDENT STIMULUS

11.1. ORAL READING COMPREHENSION

ORIGINAL (HAUSA)

Text to be read by students

Wata rana Musa da abokinsa Ali suka haɗu don su ci shinkafa. Musa ya yi loman haɗama/zarin loma, sai shinkafa ta shaƙe/sarƙe shi. Sai ya fara tari. Ali ya damu ƙwarai. Sai ya yi Sauri ya kawo mashi ruwa ya sha. Bayan Musa ya sha ruwa, sai suka gama cin shinkafarsu. Sai suka ruga a guje wajen yin wasar ƙwallo.

Questions [Answers]

1. Me Musa da Ali suka yi tare? [Sun ci abinci; sun ci shinkafa]
2. Me ya faru lokacin da suke cin abinci? [Shinkafa ta shaƙe/sarƙe Musa; Musa ya yi zarin loma]
3. Don me Ali ya damu ƙwarai? [Shinkafa ta sarƙe Musa ; Musa yana tari; Mai yiwuwa Musa ba zai iya numfashi ba; Zai yiwu ya mutu]
4. Yaya Ali ya taimaki Musa? [Ya kawo mishi ruwa ya sha]
5. Yaya kika/ka san cewa Musa ya ji sauki? [Sabo da Musa yana numfashi sosai; Ya samu ya gama cin shinkafa; Musa da Ali sun tafi yin wasar ƙwallo]

TRANSLATION

Text to be read by students

One day Moses and his friend Ali got together to eat rice. Moses took a big bite, and the rice choked him. Then he started coughing. Ali was very worried. He hurried to get him a drink. After drinking water, they finished eating their rice. They ran to play soccer.

Questions [Answers]

1. What did Moses and Ali do together? [They eat food; they ate rice] – LEVEL O1-QUESTION
2. What happens when they eat? [Rice chokes Moses; Moses had problems eating] – LEVEL O2-QUESTION
3. Why is Ali so worried? [Rice chokes Moses; Moses is coughing; Moses may not be able to breathe; He may have died] – LEVEL O3-QUESTION
4. How did Ali help Moses? [He brings a drink] – LEVEL O2-QUESTION
5. How do you know that Moses felt better? [Moses is breathing deeply again; He got to finish eating rice; Musa and Ali went to play soccer] – LEVEL O3-QUESTION

DESCRIPTION OF QUESTION LEVELS FOR ORAL READING COMPREHENSION:

LEVEL O1: Retrieve explicit pieces of information by direct word matching
LEVEL O2: Retrieve explicit pieces of information from a single sentence
LEVEL O3: Retrieve explicit pieces of information across more than one sentence

11.2. AURAL COMPREHENSION

ORIGINAL (HAUSA)

Text read to students

Zainab tana son zuwa makaranta. Kowace rana tana yin aikin gida da wani fensir da take so sosai. Watarana Zainab ta manta ba ta ɓoye fensirinta ba. ƙanenta Khalid ya ɗauka, kuma ya karya shi. Lokacin da Zainab ta zo yin aikin gida, sai ba ta ga fensirinta ba. Ta yi ta nema, ta yi ta nema. Khalid ya ga ran Zainab ya ɓaci. Sai ya fara ɗaukar ruwa ana biyansa, sannan ya saya wa Zainab wani fensir. Dukansu suka ji daɗi.

Questions [Answers]

1. Me Zainab take son yi? [Zuwa makaranta; yin karatu da fensirin da tafi so.]
2. Me ya sa ran Zainab ya ɓaci? [Ba ta ga fensirinta ba; Khalid, ƙanenta, ya ɗauki fensirinta, kuma ya karya shi; fenisirin da ta fi so ne]
3. Yaya Khalid ya ji game da karya fensir? [Ransa ya ɓaci; tsoron cewa Zainab za ta yi fushi sosai; kunya ta kama shi, kuma ya so ya sayo wani fensir; ko wata amsa da ta dace]
4. Yaya Khalid ya sami kuɗin sayen wani fensir? [ɗaukar ruwa]
5. Me ya kamata Zainab ta yi don kada haka ta faru ga fensirinta a gaba? [Ta sa fensirin a cikin jakarta bayan ta gama amfani da shi; ta hana wa qanenta ɗaukarsa; ta bar ƙanenta ya yi amfani da shi kawai in tana gida; ko wata amsa da ta dace]

TRANSLATION

Text read to students

Zainab likes to go to school. Each day she does her homework with her favorite pencil. One day Zainab forgot to hide her pencil. Her younger brother Khalid took it, and broke it. When Zainab came to do her homework, she didn't find her pencil. She searched and searched. Khalid saw Zainab was sad. Then he began to draw water to pay Zainab a pencil. They all were very happy.

Questions [Answers]

1. What does Zainab like? [Going to school; writing with her favorite pencil] – LEVEL A1-QUESTION
2. Why is Zainab so sad? [She didn't find her pencil; Khalid her brother took her pencil, and broke it; it is her favorite pencil] – LEVEL A2-QUESTION
3. How did Khalid feel about breaking the pencil? [His heart was broken; afraid that Zainab would be furious; he was embarrassed, and wanted to buy a pencil; or any other appropriate answer] – LEVEL A3-QUESTION
4. How did Khalid get the money to buy a pencil? [Draw Water] – LEVEL A1-QUESTION
5. What should Zaynab do to prevent this from happening to her pencil in the future? [Put the pencil in her purse after she finishes using it; forbid her brother to take it; let her brother use it only when she is at home; or any other appropriate answer] – LEVEL A3-QUESTION

DESCRIPTION OF QUESTION LEVELS FOR AURAL COMPREHENSION:

LEVEL A1: Identify simple inferences within single sentences
LEVEL A2: Identify simple inferences across consecutive sentences
LEVEL A3: Identify simple inferences by connecting information across the text