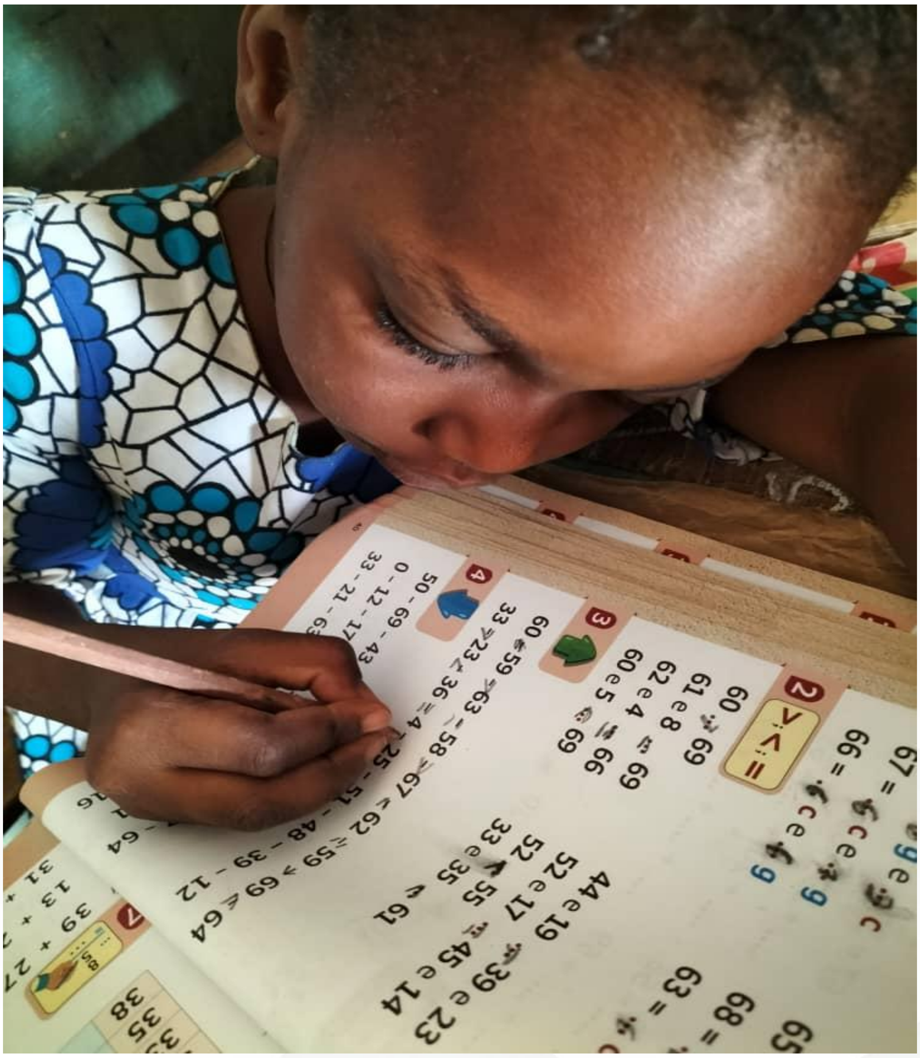
Evaluation of the *Ndaw Wune* Remedial Education Program

Final Report

American Institutes for Research, with support from the Gates Foundation

MARCH 2026



AIR®
Advancing Evidence.
Improving Lives.

Contents

Exhibits

Acronym List

AIR	American Institutes for Research
ANOVA	Analysis of Variance
ARED	Associates in Research and Education for Development
ASER	Annual Status of Education Report
DID	Difference-in-differences
EGMA	Early Grade Math Assessment
EGRA	Early Grade Reading Assessment
FCFA	Central African Franc
GDP	Gross Domestic Product
IEF	*Inspection de l'Éducation et de la Formation*
IRB	Institutional Review Board
IRR	Inter-rater Reliability
ITT	Intention to Treat
LOI	Language of Instruction
LMICs	Low- and Middle-Income Countries
MEN	Ministry of National Education
MOHEBS	*Modèle harmonisé d'enseignement bilingue au Sénégal*
RCM	Resource Cost Model
RCT	Randomized Control Trial
TaRL	Teaching at the Right Level
TOR	Terms of Reference
USD	United States Dollar

Acknowledgments

The authors would like to thank the Gates Foundation for funding this study. We would especially like to thank Clio Dintilhac and Penelope Bender at the Gates Foundation for their valuable comments and feedback on the design of the study, the data collection instruments, the methodology, and the report, and for sharing information about the context in Senegal. This evaluation would also not have been possible without the engagement of Associates in Research and Education for Development (ARED), who provided information about the *Ndaw Wune* program and its administration, including all cost data. A special thanks to our consultants Professor Mangary Ka and to Alioune Badara Diop, former Director of Institut National d'Étude et d'Action pour le Développement de l'Éducation (INEADE), for their constant support and availability. We would like to acknowledge Dalberg Research, who served as our data collection partner for this evaluation. The staff at Dalberg effectively translated and programmed the data collection tools, carried out all student assessments, managed logistics, and cleaned data within a tight project timeline. We would like to thank the many students, teachers, and school directors who cooperated with this evaluation to allow the student assessments to take place. It is our hope that insights gathered from the information they provided will help to enhance the strength of the *Ndaw Wune* program and education in Senegal overall.

Executive Summary

Senegal faces persistent challenges in foundational literacy and numeracy, with only 13% of students achieving minimum proficiency in reading and 34% in mathematics by the end of primary school *(Monitoring Impacts on Learning Outcomes,* n.d.). Despite improvements in primary completion rates, from 17% in 2000 to 51% in 2020, progress has stagnated since 2015, and disparities remain, especially among rural and low-income populations. Teachers often lack mastery of how to teach the skills-based curriculum, and many have not received the language arts teacher guide or have received only a digital version (*All Children in Senegal Are Born to Learn*, n.d.). Further, Senegal's education system has long grappled with the challenge of language of instruction. Even though almost all Senegalese children speak a Senegalese national language and do not speak French prior to starting school, the dominant language of instruction remains French, a legacy of colonial rule.

Since the 1970s, Senegal has piloted a range of approaches to bilingual education, and the 2002 education law recognized national languages as potential mediums of instruction. More recently, the *Modèle harmonisé d'enseignement bilingue au Sénégal* (MOHEBS) was introduced to institutionalize bilingual education in primary schools, aiming for national coverage by 2028. Bilingual language arts instruction is currently being implemented in 12 of Senegal's 14 administrative regions, and the implementation of bilingual math instruction started in seven regions during the 2025–2026 school year. Despite policy reforms, the use of national languages in classrooms remains inconsistent. Teachers often lack sufficient training in bilingual pedagogy, and instructional materials in national languages are unevenly available (Kébé, n.d.; Nakamura et al., 2024).

In this context, Associates in Research and Education for Development (ARED) is implementing *Ndaw Wune* ("Success for All"), a structured remedial education program that aims to support struggling CP (*cours préparatoire*, or Grade 2) and CE1 (*cours élémentaire première année*, or Grade 3) students in literacy and numeracy. *Ndaw Wune* centers around instruction in national languages and the use of leveled learning groups with targeted lessons based on an assessment of student skills. Within the *Ndaw Wune* model, teachers provide direct instruction to one skill-based student group at a time, while the other groups engage in independent work in leveled workbooks. With support from Echidna Giving, an anonymous family foundation, and the Gates Foundation, and in partnership with Senegal's Ministry of National Education (*Ministère de l'Education Nationale*, MEN), ARED has developed two versions of the program: (a) *Ndaw Wune Année Scolaire,* which operates during the school year on Monday, Wednesday, and Friday afternoons, and (b) *Ndaw Wune Vacances*, which takes place during the long July–October vacation. The MEN is currently piloting a third remediation program, *Remédiation Harmonisée,* which uses some of the *Ndaw Wune* principles to provide 30 minutes to one hour of math and 30 minutes to one hour of language arts instruction, to all students during the four hours per week designated for remedial instruction (on Tuesday and Thursday afternoons).

To provide critical evidence to inform the implementation and potential scale-up of *Ndaw Wune*, the American Institutes for Research® (AIR®) conducted an independent, quasi-experimental, longitudinal evaluation with two rounds of data collection to determine the impact of *Ndaw Wune Année Scolaire* on foundational learning outcomes. The impact evaluation used a mixed-methods design and included an assessment of the costs and the cost-effectiveness of the *Ndaw Wune Année Scolaire* program.

Evaluation Design

DID Analysis With Matching

To determine the impact of the *Ndaw Wune Année Scolaire* program, we used a difference-in-differences (DID) combined with a matching approach. We examined the impacts of *Ndaw Wune Année Scolaire* for the 2024–2025 school year cohort, which began instruction in December 2024 and ended in May 2025. We created statistically similar groups of comparison schools, using propensity score matching. By constructing statistically similar groups at the school level and then accounting for the differences between the two groups at the beginning of the program and any change in student abilities over time to a "business as usual" scenario, as captured by the comparison group, we estimate the causal impacts of *Ndaw Wune Année Scolaire*.

Our two-step methodological approach is outlined in **Exhibit 1**. In the first step, ARED selected *Ndaw Wune* treatment schools for the 2024–2025 academic year. Within the four regions (Diourbel, Kaolack, Matam, and Saint-Louis) where treatment schools were located, we selected "comparison schools" based on geographical proximity and nearest neighbor matching. Ideally, we would also have matched on observable school characteristics and language of instruction; however, that information was not available prior to baseline for comparison schools. This resulted in an imbalance in schools by language.[1] The matching resulted in 35 pairs of treatment/comparison schools that were the closest matches, for a total sample of 70 schools.

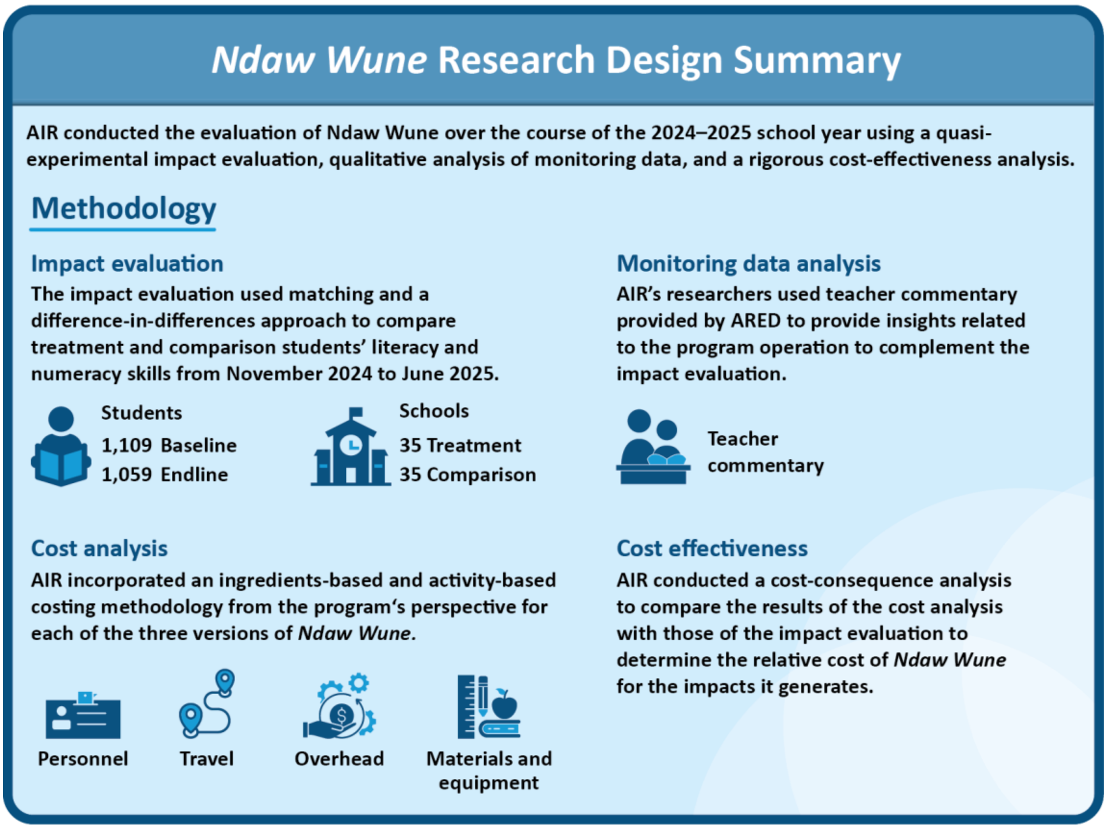
In the second step, after determining the list of matched treatment and comparison schools, AIR selected a group of comparison students to mimic the selection of students in the treatment schools. In treatment schools, ARED administered a literacy and numeracy assessment to CE1 and CP students. Those who scored below a pre-determined threshold (e.g., the ability to read at least 10 simple, familiar words) were considered eligible for the program. From the eligible students, we randomly selected 16 students per treatment school to create a sample of treatment students. To ensure that comparison students in our sample closely aligned with those selected for remediation in *Ndaw Wune* schools, we followed ARED's process and administered the same selection test to students in the comparison schools. After determining the 35 eligible students at each comparison school, we randomly selected 16 to form the sample of comparison students.

Once comparison students and treatment students were identified, we collected baseline data at the start of the school year (November–December 2024) and follow-up data at the end of the school year (May–June 2025). At endline, we were unable to find 23 percent of the sample, however, the differential attrition between the treatment and comparison groups was low, at 5 percent. We sampled replacement students at endline (using similar criteria as noted above) to ensure we had a large enough sample at endline to detect

[1] Specifically, there was just one Serer school in the comparison group compared to 12 in the treatment group. Additionally, there were 23 Wolof schools in the comparison group compared to 12 in the treatment group and 11 Pulaar schools in the comparison and treatment groups.

meaningful effects. As a note, we present robustness checks with only those students who were at baseline and endline and the full sample including replacement students.

Exhibit 1. Methodological Approach



Costing Analysis

AIR also conducted a retrospective programmatic cost analysis of the *Ndaw Wune* remedial education program. That is, we collected costs from the program perspective, including all costs directly incurred by the program implementer. Our team produced cost estimates for the 2024–2025 school year version of the program: *Ndaw Wune Année Scolaire.* We used costs related to the regions, number of students, and schools that the program served during that time frame.

AIR adopted a hybrid approach to cost analysis, incorporating both ingredients-based and activity-based costing methods. The ingredients approach is a resource-intensive best practice methodology (Danks & Kendziora, 2021; Levin et al., 2018), which involves a systematic process to identify and quantify every program input (such as number of person-hours by role and activity, type and number of materials used, etc.) and then assigns a unit price to each input to calculate the input cost. To optimize cost efficiency and minimize data collection burden among staff and participants, we limited our data collection for the cost study to programmatic and financial documents that ARED provided to us and interviews with key ARED staff. Using the data provided, we followed the ingredients-based approach for resources for which we were able to obtain detailed unit price and quantity data, such as the number of student workbooks and travel reimbursement costs for teachers attending training.

For program inputs lacking ingredient-level data, we supplemented with an activity-based approach. The activity-based approach involves the identification of necessary program inputs for each program activity and the subsequent allocation of value, using program budgets, expenditure documents, activity records, and key informant interviews (Brander et al., 2024). That is, instead of gathering data about the number of hours each ARED staff spent on each activity within *Ndaw Wune* and then multiplying that quantity by the labor price per hour to determine personnel costs for each program component, the activity-based approach uses program activity descriptions, annual salary data, and interviews with ARED organizational staff to proportionally allocate an estimated personnel cost to each program activity. This hybrid approach is an efficient methodology to produce rigorous and usable cost data that can inform program scale-up.

Qualitative Analysis of Monitoring Data

To supplement the evaluation, ARED shared qualitative data from their monitoring system for all treatment schools. This data included brief commentary from 54 teachers from December 2024 through April 2025. Teachers in *Ndaw Wune* are government teachers who are paid a supplemental stipend to implement the program. Senegal has many different categories of teachers; *Ndaw Wune* teachers include civil servants, contract teachers, and community teachers, all of whom have different levels of training, experience, and renumeration. AIR's researchers used the teacher commentary to provide insights related to the program that could be helpful as ARED and the MEN to consider while scaling up the program. Our team followed a systematic process to qualitatively analyze the teacher feedback. Using NVivo qualitative research software, we first uploaded the cumulative commentary and classified the data by (a) supervisor, (b) region, and (c) school. We then inductively coded the commentary according to emergent themes such as "student absences" or "lack of materials." After completing the thematic coding process, we used the findings to provide insights related to program implementation. However, as a note, since the teacher commentary data is self-reported and anecdotal, it is only used to provide suggestions of themes which might be helpful to consider as the program scales up and not necessarily to explain any impact evaluation findings.

Results

Impact Evaluation Results

Our evaluation is informed by the theoretical principles of the Simple View of Reading (Gough & Hoover, 1989), which is more recently conceptualized as the Cognitive Foundations of Reading Acquisition (Hoover & Tunmer, 2020). Within this framework, basic reading comprehension is contingent on the integration of language proficiency and processing skills and word recognition (or decoding).

Overall, the analysis shows that *Ndaw Wune* resulted in a positive impact on foundational literacy skills. Treatment students had significant and positive results in several literacy categories, including decoding skills (0.42 *SD*) [2], oral reading fluency (0.32 SD)[3], and reading

[2] Decoding is a construct of three decoding measures. Treatment student letter sound knowledge increased from 70.75 out of 100 at baseline to 83.52 at endline, invented word decoding increased from 5.99 out of 50 at baseline to 11.85 at endline, and familiar word decoding increased from 6.74 out of 50 at baseline to 12.87 at endline.

[3] Oral reading fluency for treatment students increased from 5.57 out of 50 at baseline to 13.12 at endline.

comprehension (0.14 *SD*)[4]. *Ndaw Wune* also had smaller, but still positive impacts on students' numeracy outcomes. Two of the six numeracy subtests showed positive and significant impacts, including number sense (0.19 *SD*)[5] and subtraction (0.20 *SD*)[6]. All other subtests had no significant finding.

We also examined heterogeneous impacts by gender, language, and region. As a note, the heterogeneous effects analysis was not powered to detect statistically significant impacts and findings are thus suggestive. Impacts were more pronounced for boys. While both boys' and girls' outcomes were positively impacted by *Ndaw Wune*, boys saw positive and significant impacts in three literacy subtests compared to two for girls, with the difference being reading comprehension. The results are more mixed for numeracy, but boys had positive treatment effects on the addition and subtraction sections while girls had a positive treatment effect on number sense and a negative one on the counting subtest.[7]

In terms of language, the data indicate that Wolof speakers generally had stronger treatment effects on the literacy subtests than Pulaar speakers.[8] Although Wolof-speaking treatment students showed a marginally negative treatment effect on listening comprehension ($p < .10$)[9], they demonstrated stronger positive gains in decoding skills (0.49 *SD* vs. 0.33 *SD* for Pulaar speakers) and achieved a 0.53 *SD* treatment effect on oral reading fluency, while the treatment effect for Pulaar speakers in oral reading fluency was not statistically significant.

Regionally, the strongest results came out of Kaolack. In Kaolack, three of the five literacy subtests (phonological awareness, decoding, and oral reading fluency) were significantly positive, compared to one for Saint-Louis and none for Matam. As noted above, since the analysis was not powered to detect effects at the subgroup level, an absence of impacts in Matam could be due to the study being underpowered. None of the individual regions had a positive and significant treatment effect in the numeracy section.

Costs and Cost-Effectiveness

We estimate the total cost of *Ndaw Wune Année Scolaire* to be 251,141,271 FCFA (433,002 USD) or 83,714 FCFA (144 USD) per student for the 2024-2025 school year. We find that most of the cost is dedicated to the direct instruction activities (62%), followed by

---

[4] Reading comprehension for treatment students increased from 0.68 out of 6 to 1.35 at endline.

[5] Number sense for treatment students increased from 4.48 out of 12 at baseline to 5.84 at endline.

[6] Subtraction for treatment students increased from 5.12 out of 12 at baseline to 7.26 at endline.

[7] At baseline, girls in the treatment group had statistically significantly higher scores than boys in the treatment group for two of the three decoding outcomes, addition and subtraction. The larger magnitude of the treatment effect for boys could be explained by boys having more room for growth than girls, relative to their baseline values. Related studies have found mixed gendered effects depending on the context. For instance, in the study of the remedial literacy program in South Africa, the authors noted that girls outperformed boys at baseline and endline and that the girls' advantage held even in their multivariate regression models (Fleisch et al., 2017). The evaluation of a remedial program in Niger did not find any gendered differences in treatment effects (Maruyama & Kurosaki, 2021).

[8] We examined the language of instruction matching the language(s) students knows as a potential explanation for Wolof students' performing better on the assessments. However, across all languages we see a high match (in the range of 96-100%) between the language of instruction and the dominant languages for children. Additionally, in terms of baseline levels, Wolof speakers had higher scores on most sub-skills than Pulaar speakers which seems to have continued into endline.

[9] At baseline, treatment students in Wolof-speaking schools had high listening comprehension test scores - 4.4 out of 6 compared to 3.4 out of 6 for comparison students. While both groups did improve between baseline and endline, a ceiling effect may have prevented the treatment students from improving at the same or greater rate than the comparison students leading to the negative impact estimate.

management, planning, and evaluation (23%). About half of program costs (51%) go to direct personnel costs, while nearly one-fifth (20%) of costs are used to fund travel for trainings, pedagogical days, and evaluations. The cost-effectiveness analysis finds that a $100 investment per student produces moderate effect size increases in letter sound knowledge, word decoding, and oral reading fluency, as well as small effect size increases in reading comprehension and numeracy outcomes.

We also calculated cost projections for four scaling scenarios for *Ndaw Wune Année Scolaire* (scaling to 50,000 students in the six IEFs):

- **Scenario 1:** No change to the number of instructional hours, initial teacher training, or pedagogical days. School directors assume the role of supervisors. The estimated cost for Scenario 1 is 767,650,752 FCFA (1,323,536 USD). The per-student cost is 15,353 FCFA (26 USD).
- **Scenario 2:** No change to the number of instructional hours. The initial teacher training is reduced to four half-days. Stipends are removed for supervisors and inspectors. The estimated cost for Scenario 2 is 638,388,975 FCFA (1,100,671 USD). The per-student cost is 12,768 FCFA (22 USD).
- **Scenario 3:** All instruction occurs during official remediation school hours. The initial teacher training is reduced to four half-days. Stipends are removed for supervisors, inspectors, and teachers. The estimated cost for Scenario 3 is 360,055,641 FCFA (620,786 USD). The per-student cost is 7,201 FCFA (12 USD).
- **Scenario 4:** All of the assumptions of Scenario 3, and students use existing materials such that there are no materials costs for printing/delivering instructional materials. The estimated cost for Scenario 4 is 263,343,808 FCFA (454,041 USD). The per-student cost is 5,267 FCFA (9 USD).

Qualitative Analysis of Monitoring Data Results

The *Ndaw Wune* program was designed to be two hours of remediation, three days a week, with about 60 minutes for literacy (including the alphabet song), 30 minutes for math, 15 minutes for math and literacy educational games, and about 15 minutes for the alphabet song, recess, and assigning homework at the end of the class. The program model divides students into three groups for literacy: one for students who can identify fewer than 10 letters, one for students who can identify at least 10 letters but fewer than 10 syllables, and another for students who can read (decode) at least 10 syllables but fewer than 10 words. It includes two groups for numeracy: one for those who can correctly complete fewer than five number comparisons and one group for students who could correctly complete all five number comparisons.

Qualitative analysis of ARED monitoring data indicated that the program provided significant support to students in several ways. Teachers noted that the provision of teacher guides, student workbooks, and alphabet flashcards helped improve lesson delivery and made learning activities more effective. The program's structure allowed for remediation and flexible movement between groups, enabling students to receive targeted support based on their progress and needs. Teachers reported some challenges to implementation. These are self-reported anecdotal information reported by teachers and were mentioned by fewer than 1-7% of teachers in each instance. Common issues included perceived large number of students, making it difficult to manage the class, high rates of student absences, missing or absent materials for instruction, inaccuracies in materials, and difficult-to-use numeracy materials.

Limitations

This study faces some limitations.

1. There was limited school-level data available for matching treatment and comparison schools. Ideally, we would have been able to match based on observable characteristics such as student and teacher attendance rates, dropout rates, existence of school meals and other school-level resources, and, importantly, the language of instruction at the school. However, much of this data was not available prior to data collection when matching was conducted. Thus, while the school-level matched sample is based on the limited variables that were available, the matches could potentially be different for many other observable and unobservable characteristics. Relatedly, since we did not have access to language of instruction information about schools before matching was done, in the eventual school sample we had data for only 12 Serer treatment schools and only one Serer comparison school (data on language of instruction was only available for comparison schools once enumerators went to schools for data collection). This was not a large enough sample to make reliable comparisons between the treatment and comparison groups for Serer and thus in this study we are unable to draw conclusions for the Serer subsample.
2. We had limited statistical power to detect heterogeneous effects. While our sample size is large, our evaluation design includes a limited number of schools, constraining our ability to detect heterogeneities in the impact estimates with sufficient precision.
3. A key assumption when using the DID methodology is that treatment and comparison students were on identical tracks before implementation began (i.e., without *Ndaw Wune*, the treatment and comparison students would have had the same trend in their outcomes). This is usually tested by collecting data from multiple points before implementation begins. It is common (as in this case) that previous data collection is not possible, and the matching strategy and baseline equivalence tests help to address this limitation.
4. For the cost analysis, while our hybrid approach of combining ingredients-based and activity-based costing is both appropriate and efficient for this analysis, there is a trade-off in terms of precision. For programming activities for which we did not have detailed expenditure data, we estimated costs based on activity descriptions, terms of reference documents (TORs), and interviews with ARED organizational personnel. Our cost data could also be subject to recall bias of program implementers. Triangulation between the documented cost data and the interview data helped to mitigate the drawbacks of recall bias.

Conclusions and Recommendations

The evaluation demonstrated that the *Ndaw Wune Année Scolaire* program yields significant improvements on key foundational literacy skills (decoding: 0.42 *SD*, oral reading fluency: 0.32 *SD*, reading comprehension: 0.14 *SD*) and modest gains in numeracy (number sense: 0.19 *SD*, subtraction: 0.20 *SD*) for struggling early grade students. These results provide robust evidence that targeted, skill-based remedial instruction, delivered in languages children know and organized by student proficiency groups, can address foundational learning challenges in Senegal. Policymakers should prioritize the scale up of such remedial models, especially in regions and populations with the lowest learning outcomes.

We present key recommendations for next steps in strengthening foundational literacy and math education in Senegal. These recommendations can be implemented through *Ndaw Wune* or through other MOHEBS programming and support.

For Teacher Professional Development and Classroom Instruction:

1. **Expand explicit phonics instruction in regular classrooms**. The *Ndaw Wune* program demonstrated that supporting grapho-phonological mapping skills through explicit phonics instruction and providing structured opportunities for improving automaticity (speed and accuracy of making sound-symbol connections at the sub-lexical and lexical levels) are proving highly effective for improving decoding sub-skills and reading comprehension. Given the strong positive results for these sub-skills from the *Ndaw Wune* evaluation, we also recommend that MOHEBS draw from these findings and consider integrating additional explicit phonics instruction in Senegalese languages in classrooms across the country, which has not been happening in most classrooms (over 73%, Nakamura et al., 2024) in Senegal.
2. **Ensure reinforced implementation of the structured pedagogical elements for teaching foundational skills** in classrooms across Senegal. The combination of scope and sequence-based planning, clear and scripted (more so for language arts than for math) teacher guides, and student and teacher learning materials, especially workbooks for each child, have significant impact, as per the results of this evaluation. These aspects of instruction should be the core of pre-service teacher preparation programs and in-service training and tracked through implementation fidelity monitoring in foundational learning programs across the country. Given large class sizes, the lack of teaching and learning materials other than textbooks and teacher guides in MOHEBS classrooms, minimal exposure to print outside of school, and chalkboards that are in poor condition in many classrooms, the importance of a leveled workbook for each child is a critical component of the *Ndaw Wune* approach.
3. **Consider further oral language support in national languages.** The results of the evaluation found no significant improvement in oral language skills – one of the necessary core components of foundation literacy development. Further, this result was not due to a ceiling effect in oral language skills at baseline. Therefore, *Ndaw Wune* and other programs nationally must bolster listening comprehension and oral vocabulary skills – even if already teaching in languages children know and understand. In other words, even when children are taught in their own language, there is a need to spend considerable teaching time on bolstering vocabulary and linguistic knowledge across the curriculum and not focus solely on phonics and decoding instruction.
4. **Invest in improvements for numeracy components of the program.** The impact evaluation revealed challenges with numeracy instruction. Therefore, we provide some suggested approaches for improving foundational numeracy instruction. First, instruction should follow clear progression within each operation, moving from

viewing multi-digit numbers as strings of individual digits to understanding the full place-value structure (e.g., 10 ones = one 10) (Rogers, 2023). Second, teachers could explicitly teach efficient calculation strategies beyond tallying, such as bridging through ten, jumping, and rounding and adjusting methods shown to significantly improve mental math fluency when modeled and practiced systematically (Kumm & Graven, 2024). Third, structured number talks could be incorporated to encourage students to share and compare multiple solution strategies, fostering number sense, flexibility, and metacognitive growth through collaborative discourse. Finally, extensive use of grouping materials, such as bundles of sticks or locally made base-ten blocks can make abstract place-value concepts tangible (e.g., reinforcing the idea that 10 ones form a ten) and support regrouping in operations (Resnik, 1987).

5. **Need for supporting transfer and French oral language skills given Senegal's bilingual education goals.** The goal of *Ndaw Wune* is improvement of literacy skills in national languages to give students the foundation they need to learn to read in a second language. The goal of foundational learning in Senegal is literacy ability in national languages and French (Nakamura et al., 2026). The science of bilingual reading makes it clear that learning to read in a new language is critically dependent on how well students can read in a first language, and how much oral (or signed) language proficiency they have in the new language they need to learn. More emphasis on the cognitive underpinnings of biliteracy acquisition would better support students' transfer of their L1 skills to an L2.

For Further Research:

1. **Understand the intersections between language and numeracy development.** Future research and program design in Senegal should systematically investigate how language skills and the choice of language of instruction shape students' numeracy development. Research should explore how bilingual or multilingual teaching, where students learn math in both their home language and French, affects mathematical reasoning, transfer of skills, and long-term achievement. For example, does early mastery of numeracy in Wolof or Pulaar facilitate later learning of math in French, or does it create challenges in transitioning mathematical vocabulary and concepts? Future research should also study how language mismatches affect numeracy outcomes over time.
2. **Ensure data on matching teacher/student language is collected.** With the advent of large-scale initiatives in national languages, the MEN plans to collect teacher language data more systematically in the future within the MIRADOR (information system for teacher management) to support the implementation of bilingual education. We recommend supplementing this with student language data to understand the extent of student-teacher language match/mismatch.
3. **Carry out a process evaluation to delve into the elements of the program that most likely lead to improved student achievement and the reasons for the regional variations.** As ARED seeks to refine and scale-up *Ndaw Wune,* a corresponding process evaluation would be tremendously useful in identifying key elements of the program that are helpful for students and teachers and noting how the program can replicate the successes and address any issues in the program to improve it for scale-

up. Such an evaluation could examine the fidelity of implementation by comparing intended activities such as structured lesson delivery and use of instructional materials with what occurred across schools. It could also assess the quality of delivery through classroom observations that capture how well teachers apply the pedagogical methods promoted by the program, the consistency of coaching support, and the nature of any adaptations made in practice. Gathering teacher perspectives through interviews, surveys, and focus groups could help illuminate how teachers perceive the usefulness of the training and materials, the challenges they face, and the conditions that most enable or hinder effective implementation. The evaluation could also look closely at student engagement and classroom dynamics to identify the instructional practices that most consistently lead to active participation and comprehension. In addition, tracking program dosage (i.e., how much exposure students have to *Ndaw Wune* lessons in practice) could help explain variation in effects across schools and regions.

4. **Systematically track student attendance data.** Since the attendance data we received was inconsistent and had several gaps, future work should explore instituting a systematic and cost-effective data tracker which can track individual student attendance. This would help in identifying reasons for absences and uncover chronic absenteeism. This can include a simple SMS-based tool that can feed into a data-driven decision-making tool and would not require internet or smartphones. Such an approach has been successfully piloted in several countries, including those in West Africa (Mugo & Piper, 2025). For instance, a pilot in Kenya found that the cost of maintaining an SMS-based attendance system for 10,000 learners at a central level, supplemented by weekly personalized SMS to teachers on struggling learners would cost less than 0.08 USD per child per year (Mugo & Piper, 2025). Improved data systems will also help track attendance, attrition, and heterogeneous impacts, enabling more responsive program management.
5. **Utilize artificial intelligence (AI) and education technologies (where appropriate) to assist with provision of additional teaching and learning materials (TLMs) and trainings.** AI-generated short videos could be used for frequent light-touch training sessions and supportive coaching. It could also be a supplementary input on tough topics or practice sessions on numeracy-related topics. It may also be useful to utilize AI to produce additional "back-up" materials in line with all the science of bilingual reading principles discussed above, with the caveat that AI translation and content generation are not yet available for Senegalese languages. Some examples include:
    a. Assurance that every student has access to textbooks/workbooks, every teacher has a guide and they are aligned. In rare cases, where students need to share books, teachers should be guided as to how to guide students through sharing activities (part of the teacher training).
    b. Utilizing AI to create pictures that can be used as story cards, letter cards, and other TLMs. These can be created much more cost-efficiently and are not always reliant on current linguistic constraints of AI.
    c. Ensuring materials within the student textbook address all literacy and language sub-skills in an appropriate learning trajectory, especially for

biliteracy acquisition and later French transition, for there to be fewer pieces to print and align.

For Policy:

1. **Importance of initial reading instruction in Senegalese languages**. The evaluation of the *Ndaw Wune* program makes it clear that when teaching children in languages they use and understand, with a focus on explicit teaching of letter sounds and phonics rules, significant improvements occur in students decoding, and reading comprehension, even for struggling, high risk students. *Ndaw Wune* is an exceptional example of how, when students are taught in languages they understand, resources can be focused on the word recognition aspect of foundational literacy. This approach, as opposed to teaching children in a foreign language, makes literacy acquisition much more efficient as precious classroom time is not wasted on teaching introductory oral language skills, and focuses on letter sounds, blending, fluency, and higher-level oral language skills.

   In addition to the use of a language students know, it is also vital for teachers to know the language of instruction. The data from this study did not find differences between student language knowledge and the language of instruction, but we did not collect data on teacher language knowledge. There are still considerable mismatches across Senegal between students, teachers and the language of instruction in particular schools, especially for languages spoken by a smaller population.[10,11] These need to be addressed for effective foundational learning and numeracy bilingual instruction.

[10] https://www.air.org/sites/default/files/2025-08/BUILD-Introduction-to-the-Framework-August-2025.pdf
[11] https://scienceofteaching.site/wp-content/uploads/2025/08/English-Implications-of-and-Strategies-for-Dealing-with-Language-of-Instruction-Mismatch-in-Senegal_Final.pdf

1. Introduction

To improve foundational learning in Senegal, the Associates in Research and Education for Development (ARED) supports bilingual education across the country. Currently, ARED serves as the government's key partner for the development and implementation of the *Modèle Harmonisé d'Enseignement Bilingue au Sénégal* (MOHEBS, Bilingual Education Framework for Senegal), providing technical assistance related to instruction in national languages and transition to French, including developing textbooks, teacher guides, and books for reading practice and supporting teacher professional development. ARED is also supporting the development of *Remédiation Harmonisée,* led by the MEN, which provides remedial instruction in math and language arts in one hundred first and second grade classrooms as part of the official school day. ARED's deep understanding of the educational and linguistic context of Senegal continues to pave the way for education reform across the country.

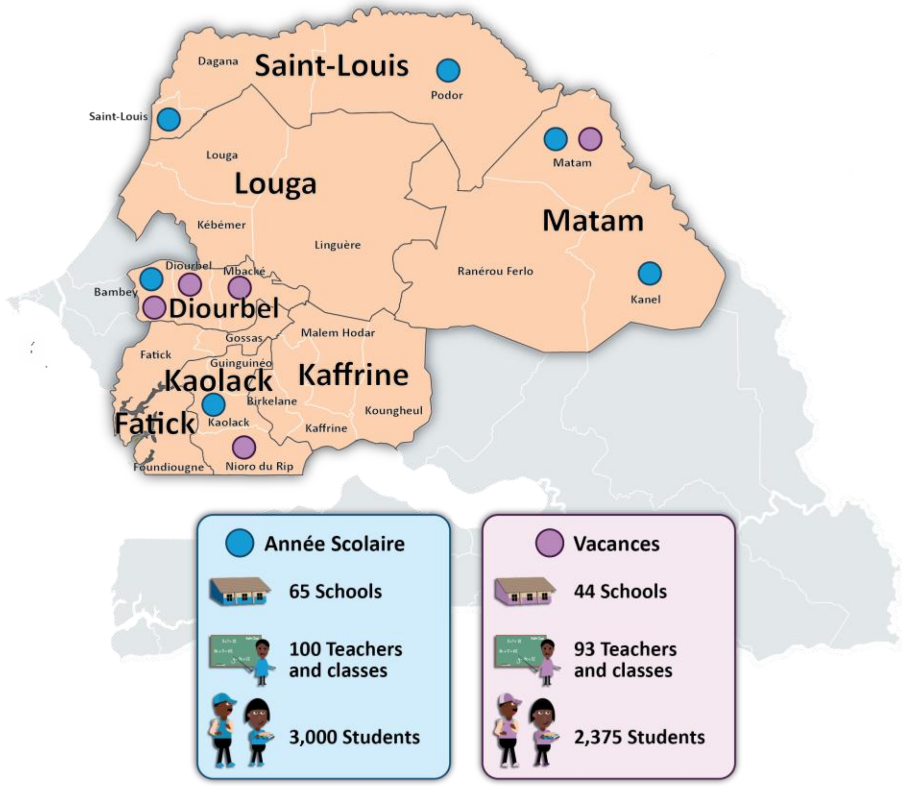
In addition to its work on ministry programs like the MOHEBS and *Remédiation Harmonisée*, ARED began developing its own smaller scale remediation program, *Ndaw Wune* (Success for All) in 2021. *Ndaw Wune* is a structured remedial intervention program that provides foundational literacy and numeracy instruction to struggling Senegalese early grade learners who are at risk of school failure and dropout using one of three largest national languages—Pulaar, Serer, or Wolof. The program mainly targets students in CE1 (*cours élémentaire première année*, or Grade 3) and includes some CP (*cours préparatoire*, or Grade 2) students. Using the test shown in **Appendix A,** the program divides students into three groups for literacy: one for students who can identify fewer than 10 letters, one for students who can identify at least 10 letters but fewer than 10 syllables, and another for students who can read at least 10 syllables but fewer than 10 words. It includes two groups for numeracy: one for those who can correctly complete fewer than five number comparisons and one group for students who can correctly complete all five number comparisons (but qualify for the program based on their reading score). Students participate in one group for literacy and one group for numeracy during the instructional time for each subject. Groups receive direct instruction once a week for literacy and 1.5 times a week for numeracy, while participating in teacher-guided, independent work in leveled workbooks during the remainder of the allocated program time.

Currently, there are two versions of the program: (a) the School Year, or *Année Scolaire,* version of the program (the focus of this report), and (b) the Vacation, or *Vacances*, version. For *Ndaw Wune Année Scolaire* version—the original iteration of the program—using the test detailed in **Appendix A**, ARED selects the lowest-performing students in CE1 and CP to participate in afterschool classes for two hours per day, three days a week (six hours per week total), for as close to 25 weeks (the official length of the school year) as possible, given constraints of delayed school start dates, holidays, weather, local events, and teacher absences. This approach features instruction in the national language (Wolof, Pulaar, or Serer) used by students in the school for classes of 30 students. In Senegal's education system, regional-level structures are referred to as IAs, which stands for *Inspection d'Académie,* and regions are divided into zones served by *Inspections d'Education et de Formation* (IEF). For the 2024–2025 school year, ARED implemented *Ndaw Wune Année*

Scolaire in one hundred classes with a total of 3,000 students in six IEFs located in four IAs: Diourbel, Kaolack, Matam, and Saint Louis.

The *Ndaw Wune Vacances* version uses a similar model, but occurs for three hours a day, three days a week, for eight weeks during the long school holiday from June through October. **Exhibit 2** provides key facts about each of the program versions for the 2024–2025 school year and 2024 vacation. A full list of program locations can be found in **Appendix B**.

Exhibit 2. Overview of *Ndaw Wune* Programming in 2024–2025



To date, no independent evaluation has determined the causal impact of the *Ndaw Wune* program on student literacy and numeracy outcomes, although there have been multiple internal evaluations and process evaluations. ***Therefore, to provide critical data to inform foundational learning programming implementation in national languages and scale-up, the American Institutes for Research (AIR) conducted an impact evaluation and cost-effectiveness analysis of Ndaw Wune Année Scolaire.*** In this report, we present the methods of our evaluation and analyses, the findings of our study, and recommendations and conclusions.

2. Theory of Change

The Cognitive Foundations of Reading Acquisition (CFRA) (Hoover & Tunmer 2020), which builds on the Simple View of Reading (Gough & Hoover, 1990) is the main theoretical underpinning for our study. This model postulates that there are two necessary components for basic reading comprehension acquisition – word recognition abilities and language comprehension. No matter what language a child learns to read first, they must necessarily be able to both sound out (decode) words accurately and quickly and be able to understand the meaning of the written word. Word recognition abilities are made up of a series of sub-skill pre-requisites such as letter knowledge skills (both an ability to name letters as well as an automatic ability to identify grapheme-phoneme correspondences) and decoding skills (i.e., the ability to sound out words quickly and accurately).

Language comprehension is also made up of a series of sub-skill pre-requisites: phonological awareness (i.e., the ability to identify, recognize, and manipulate sound [phonological] segments within one's oral language), oral vocabulary, syntax, and background knowledge to comprehend oral language. This makes it clear that the acquisition of reading comprehension is impossible in a language a child does not understand. This is not only intuitive but backed by research (Ouellette & Beers, 2010).

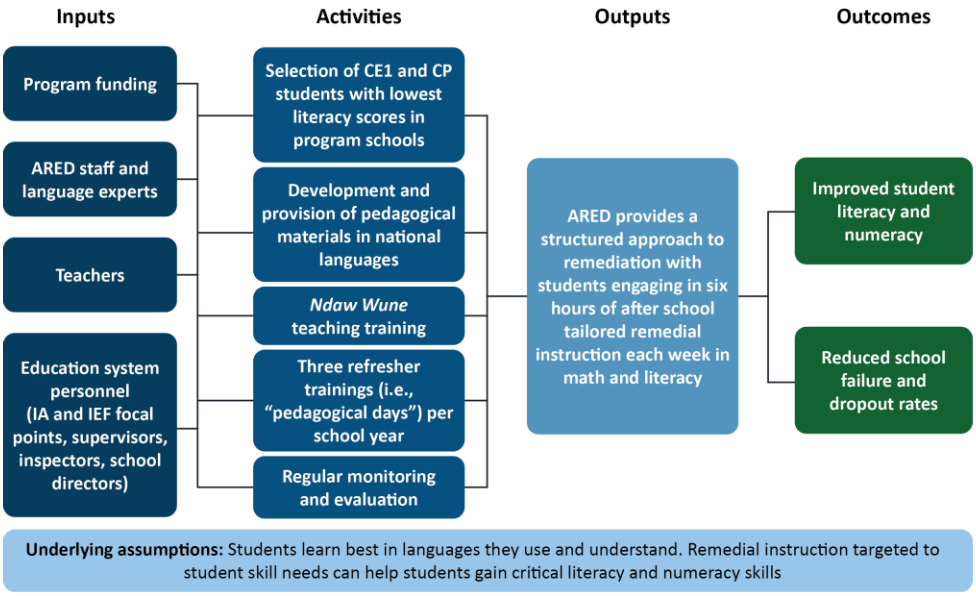
Importantly, *teaching effectiveness* in reading is tied to curriculum, teacher training models, and TLMs that align with these theories (Hoover & Tunmer, 2020). In other words, curriculum, materials, and teaching methods that explicitly teach word recognition and teach children in language(s) they understand and speak well are much more likely to succeed in fostering literacy skills than those that are not effectively doing so. Yet, in Senegal, at least 60% (and by some estimates more than 73%) of children across the country are taught in a language they do not understand (Nakamura & Holla, 2024; Nakamura et al., 2024).

Drawing on this theoretical basis for teaching children to read in a **language they know**, we provide a theory of change as a valuable tool for understanding how *Ndaw Wune Année Scolaire*, which focuses on a structured approach to remediation, works to improve foundational learning outcomes.[12] In **Exhibit 3**, we provide a basic theory of change for *Ndaw Wune Année Scolaire*. In this case, given the timeframe, we focus on the immediate outcome of improved student literacy and numeracy, rather than the medium-term effects of reduced repetition and dropout or the long-term effects of improved academic outcomes on students and their communities.

The structured remediation approach of *Ndaw Wune* is predicated on two key assumptions: (a) students learn best in languages they use and understand, and (b) remedial instruction that is targeted to the skill needs of students can help them gain critical literacy and numeracy skills. Evidence has widely demonstrated the benefits of teaching in the languages students use and understand (Nag et al., 2019; Nakamura et al., 2024; Nishanthi, 2020), including better learning achievement outcomes in both children's familiar languages and targeted international languages (Evans & Acosta, 2021; Koda, 2008; Mohohlwanea et al., 2024; Stern et al., 2024). Similarly, remedial education, including through structured learning methods and added instructional time, has been shown to improve student learning outcomes (Moussa et al., 2021; Piper & Dubeck, 2024; Snilstveit et al., 2015). ARED designed the *Ndaw Wune* model based on these theoretical strong points and existing evidence.

[12] We do not draw on second language reading given the focus of the program on fostering literacy skills in children's first languages – the national languages of Senegal.

Exhibit 3. Theory of Change for *Ndaw Wune Année Scolaire*



Note: This study only examines student literacy and numeracy as the main outcomes and does not examine school failure and dropout rates.

3. Background

The Government of Senegal invests substantially in the education sector, allocating approximately 23% of its national budget and 6% of its gross domestic product (GDP) to education in 2023 (UNESCO, 2024). Despite these significant financial commitments, persistent challenges remain: 24% of primary school-aged children are out of school, and 69% of primary school-aged students lack basic reading skills (World Bank, 2024). Further, only 18% of students aged nine to 16 demonstrate adequate reading proficiency, while just 22% are proficient in mathematics (LARTES-IFAN, 2017). According to the 2019 PASEC report, only 30% of Senegalese students are proficient at PASEC Level 4, defined as having the ability to understand narrative and informational texts, interpret implicit meanings, and understand author intentions. Factors such as teacher quality, cost to attend school, and availability of supplies, have been found to be correlated with dropout rates (Ndour, 2024).

In 2017, with the goal of increasing student reading outcomes in the earliest grades, Senegal implemented the USAID-funded *Lecture Pour Tous* (LPT) ("reading for all") program. By the end of 2021, the program had reached more than 600,000 students in Grades 1 through 3 from 3,700 public primary schools in six regions (Bagby et al., 2022). LPT was succeeded by the *Renforcement de la Lecture Initiale pour Tous* program (RELIT) an early grade language arts program, which formed part of Senegal's national reform of bilingual education until it closed along with USAID in early 2025. Under RELIT, full-scale rollout of the language arts component of the national bilingual education program through USAID government-to-government (G2G) assistance was expanded from one region, where GtG assistance was successfully piloted under LPT (Saint Louis) to six additional regions: Diourbel, Fatick, Kaolack, Kaffrine, Louga and, Matam (USAID, 2022). Direct implementation by RTI, the RELIT implementing partner, was conducted in two regions, Kedougou and Tambacounda. While LPT was proven to be effective, most students were not reading fluently by Grade 3, which indicated that structured remediation approaches might be needed to supplement classroom instruction.

In this context, in Senegal and other low- and middle-income countries (LMICs), remedial education programs offer a complementary approach to improving instruction in the main classroom and have been shown as being particularly effective in addressing foundational learning gaps in early grades (Singh et al., 2024; Moussa et al., 2021; Brown et al., 2022; Duflo et al., 2020). Programs such as Pratham's Teaching at the Right Level (TaRL) have demonstrated substantial improvements in reading fluency and comprehension among primary school students, and structured interventions focusing on number sense, counting, and basic operations have led to measurable improvements in math proficiency (Duflo et al., 2024; Angrist et al., 2023; Maruyama et al., 2022). A large-scale systematic review by the International Initiative for Impact Evaluation (3ie) analyzed four remedial education programs in Chile, India, and Mexico, and consistently found improved learning outcomes, especially when combined with structured pedagogy and additional instructional time (Snilstveit et al., 2016). All four programs targeted students who were identified as lagging behind their peers and provided tailored teaching in core skills, typically to groups that were smaller than the usual class. The teachers in all four programs were either volunteers or hired from the local community and were not trained teachers.

Ndaw Wune combines key elements of structured pedagogy: instruction that relies on carefully designed language scope and sequences for the key theoretical underpinnings of reading instruction, varying scripted lesson plans, and accompanying student workbooks. These elements have been shown in the literature to work together to improve student learning (Piper et al., 2021). There is ample evidence in support of structured pedagogy programs in LMICs, defined most often as leveled learning groups, providing scripted lessons to teachers for every day of the program (with reduced scripting as the weeks progress), and corresponding sequenced activities for students in workbooks. These programs also generally provide teacher training and coaching aligned with the preceding principles. Several studies have been conducted on small-scale structured pedagogy programs, and more recently evidence has started emerging about larger scale projects (Piper & Dubeck, 2024). For example, in a 2020 meta-analysis, Evans and Acosta found that structured pedagogy had positive results for literacy and numeracy in most of the studies analyzed in sub-Saharan Africa. Another meta-analysis found that structured pedagogy programs had the "largest and most consistent positive average effects on learning outcomes" with an overall effect of 0.23 SD (Snilstveit et al., 2017). An additional meta-analysis in LMICs showed that structured programs had an overall effect of 0.30 SD on foundational literacy outcomes, with stronger effects on lower order word recognition sub-skills, and with impact differing with degree of teacher support and on the whether the program was in an first language (L1) or second language (L2) (Kim, Lee, & Zuilkowski, 2020). A study that looked at common themes between LMIC programs that were successfully scaled found that six of the eight programs were structured literacy programs that generally included structured learning materials for students, teacher guides, and focused training for teachers alongside ongoing coaching (Stern et al., 2021). In 2023, the Global Education Evidence Advisory Panel listed structured pedagogy as a "great buy."

However, while there is ample evidence on the impact of structured pedagogy programs, studies examining the causal impacts and cost-effectiveness of large-scale remediation programs which incorporate a structured pedagogy approach from sub-Saharan Africa are scarce. Further, impacts of remediation programs delivered by trained teachers in school or after school are also very limited. In Niger, a study on a remedial primary school program focused on provision of math textbooks found an impact of 0.36 SD for fourth- through sixth-grade students (Maruyama & Kurosaki, 2021). A randomized controlled trial in South Africa measuring the impact of a remedial program for fourth- through sixth-grade students learning English as their second language found null results for the full sample but moderately positive results for students with initially stronger English proficiency (0.12 SD) (Fleisch et al., 2017).

In this context, drawing from theory on reading development globally in first language contexts, and the evidence on structured pedagogy programs in LMICs, this evaluation will examine the causal impacts of a structured remedial foundational learning program in Senegal and thereby contribute to the evidence base around the effectiveness of remediation programs for struggling students with skills far below grade level in low resource LMIC classrooms.

4. Research Questions

AIR's evaluation of the *Ndaw Wune* program addresses the following key research questions:

- RQ1. To what extent is there an impact of *Ndaw Wune Année Scolaire* on foundational literacy and numeracy outcomes for participating CP and CE1 students?
- RQ2. What is the overall cost and the cost per student?
  - RQ2A. How do costs vary by program component (e.g., teacher training, instructional delivery)?
  - RQ2B. How do costs vary by expense type (e.g., personnel, materials)?
  - RQ2C. Do the program costs vary significantly by location, and, if so, what are the estimated costs by region?
- RQ3. How do costs differ based on program scaling scenario (e.g., costs to expand to 50,000 students and potential cost savings associated with modified training schedules and/or reduced personnel expenditures)?
- RQ4. If the *Année Scolaire* version of the program is effective, what is the cost per unit of effect?

Given that the CFRA theory and multiple studies have made it clear that students cannot learn to read in a language they do not understand (assuming that reading is defined as comprehending and not simply decoding, which also relies heavily on oral language knowledge), we hypothesize that *Ndaw Wune Année Scolaire* will increase foundational literacy and numeracy sub-skills for students as they are being taught in a language they know through structured pedagogy, leveled learning groups, smaller class sizes, and leveled workbooks.

The impact evaluation responds to Question 1, whereas the cost analysis responds to Questions 2 and 3. Both the impact evaluation and cost analysis contribute to responding to

Question 4 in the form of a cost-effectiveness analysis. We describe our methods for each component of the evaluation as follows.

5. Methods

5.1. Impact Evaluation

We used a difference-in-differences (DID) technique combined with a matching approach to analyze the impact of *Ndaw Wune Année Scolaire* on the literacy and numeracy skills of participating CP and CE1 students. Our specific methodology is described in this section.

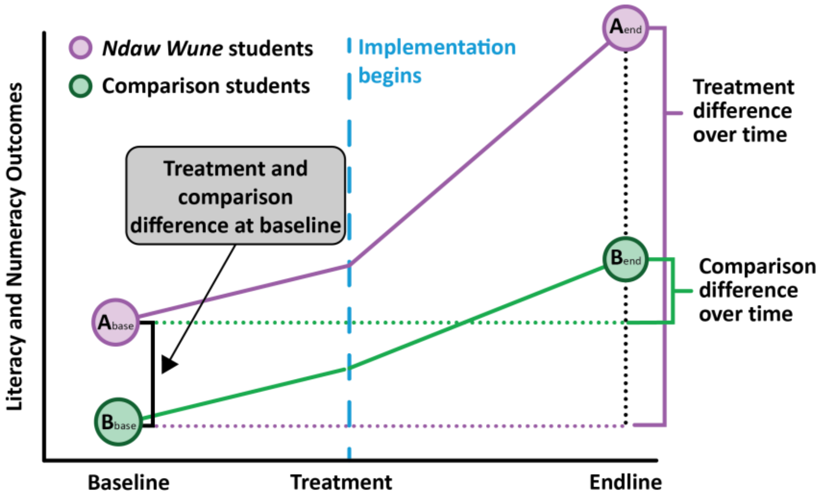
Design

We use a DID combined with a matching approach to estimate the impact of the program one school year after implementation. **Exhibit 4** illustrates this method. Specifically, we examined the impacts of *Ndaw Wune Année Scolaire* for the 2024–2025 school year cohort, which began instruction in December 2024 and ran for approximately 25 weeks. This involved first creating statistically similar groups of comparison schools using the matching approach described in detail below (Step 1) and then sampling comparison students in a similar manner as treatment students (Step 2). By constructing statistically similar groups at the school level and then accounting for the differences between the two groups at the beginning of the program and any change in student abilities over time to a "business as usual" scenario, as captured by the comparison group, we estimate the causal impacts of *Ndaw Wune Année Scolaire*.

Step 1: Matching at the school level to identify comparison schools. ARED selected 65 *Ndaw Wune* treatment schools for the 2024–2025 academic year across four regions (Diourbel, Kaolack, Matam, and Saint-Louis). Within the regions where treatment schools were located, we selected "comparison schools" by matching based on geographical proximity and school-level characteristics. We received a list of 22,539 potential comparison schools from ARED. We geographically mapped these schools; out of those, 4,438 were in the same regions (Diourbel, Kaolack, Matam, and Saint-Louis) as the treatment schools. Additionally, we received school-level data on department, urban/rural classification, total boys in school in 2023, and total girls in school in 2023 for all treatment and comparison schools. Based on the school-level data, we first estimated a propensity score (i.e., the probability of a school belonging to the treatment group, conditional on the set of observable characteristics) using a logit regression. Schools were then matched by minimizing the difference in their predicted probabilities, with each treated school paired to the most similar comparison school using nearest neighbor matching. The matching resulted in 35 pairs of treatment/comparison schools that were the closest matches, for a total of 70 schools (see **Exhibit 5** for the location of treatment and comparison schools).[13]

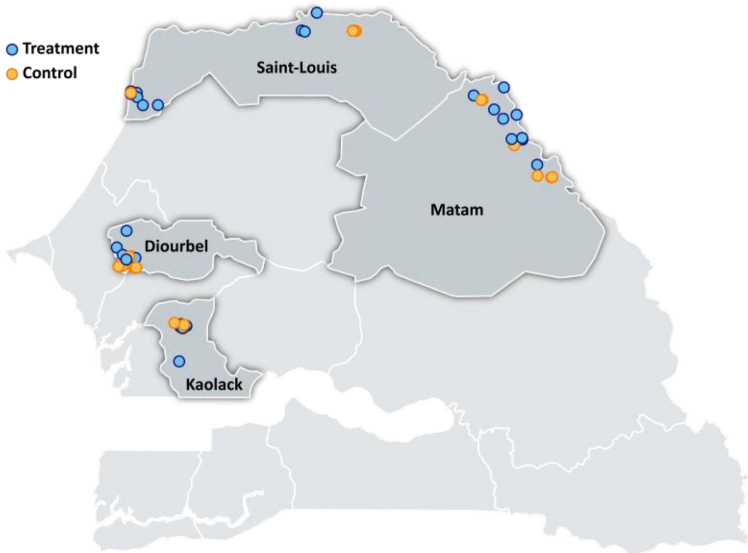
[13] Based on the following assumptions: With a sample size of 70 schools (35 treatment and 35 comparison) and 16 students per school, we would be able to detect an effect of 0.22 standard deviations (*SD*), assuming a power level of 80%; an intra-cluster correlation of 0.15 for level 2 (schools); and an R-squared of 0.40 for level 1 (students) and 0.40 for level 2 (schools). The study was only powered to detect effects amongst the entire sample and not different subgroups, and so we only provide suggestive evidence of subgroup effects. Calculations were performed using the PowerUp! Tool. Effect sizes for remediation programs have ranged from 0.06–0.27 *SD* on math and reading outcomes (Álvarez Marinelli et al., 2024; Jacob & Lefgren, 2002; Özek, 2021).

Exhibit 4. DID Approach



	Ndaw Wune students	Comparison students	Difference
Baseline	A_{base}	B_{base}	$A_{base} - B_{base}$
Endline	A_{end}	B_{end}	$A_{end} - B_{end}$
Difference	$A_{end} - A_{base}$	$B_{end} - B_{base}$	$(A_{end} - A_{base}) - (B_{end} - B_{base})$

Exhibit 5. Locations of Treatment and Comparison Schools



Even though the school-level data for matching treatment and comparison schools was limited, as shown by **Exhibits 6** and **7**, the matching resulted in schools that were mostly statistically similar to each other at baseline. Specifically, **Exhibit 6** examines balance for variables that were included in the matching, and **Exhibit 7** examines ex-post school-level balance based on the baseline data that were collected. We control for variables, which were significantly different at baseline in our final analysis.

Exhibit 6. School-Level Balance (Variables That Were Matched)

Variable	Comparison schools mean	Treatment schools mean	Pairwise *t*-test *p* value
Bambey	14%	34%	0.052*
Kanel	31%	26%	0.603
Kaolack	3%	9%	0.310
Matam	23%	11%	0.210
Podor	9%	11%	0.695
Saint-Louis	20%	9%	0.177
Urban	34%	34%	0.803
Number of boys in school	115	98	0.359
Number of girls in school	14%	34%	0.052*

Note. Standard errors in parentheses. *N* = 35 comparison and 35 treatment schools.
**p* < .10. ***p* < .05. ****p* < .01.

Exhibit 7. School-Level Baseline Equivalence (Variables Collected at Baseline)

Variable	Comparison schools mean	Treatment schools mean	Pairwise *t*-test *p* value
Electricity	86%	83%	0.747
Clean water	94%	86%	0.238
Latrine	83%	66%	0.104
Wash station	37%	14%	0.029**
Soap	20%	3%	0.025**
Free meals	11%	3%	0.168

Note. Standard errors in parentheses. *N* = 35 treatment and 35 comparison schools with the exception of one treatment school where the soap question was not able to be observed.
**p* < .10. ***p* < .05. ****p* < .01.

Step 2: Selection of treatment and comparison students. After determining the list of matched treatment and comparison schools, AIR selected a group of comparison students to mimic the selection of students in the treatment schools. To ensure comparability, we mirrored the approach that ARED used to select eligible students for the program.

Within the treatment schools, ARED administered a basic literacy and numeracy assessment to CE1 students (see sample in **Appendix A**). Those who scored below a predetermined threshold (e.g., the ability to read at least 10 words) were considered eligible for the program. If more than 35 CE1 students were eligible for the program, the lowest-scoring students were selected. If fewer than 35 CE1 students were eligible for the program, then

ARED administered the same assessment to CP students to fill the remaining spots. After ARED finalized its list of program students, we obtained the list from them and randomly selected 16 students per treatment school to form the sample of treatment students.

To ensure that comparison students in our sample closely aligned with those selected for remediation in *Ndaw Wune* schools, we followed ARED's process and administered the same selection test to students in the comparison schools, using the same process to prioritize CE1 students. After determining the 35 eligible students at each school, we randomly selected 16 to form the sample of comparison students. **Exhibit D1** in **Appendix D** details the distribution of scores for students in the treatment and comparison schools from the screening test.[14]

Step 3. Collecting baseline and endline data for the DID analysis. Once comparison students and treatment students were identified from Step 2, we collected quantitative baseline data at the start of the school year (November–December 2024) and follow-up data at the end of the school year (May–June 2025). Using the DID approach, we contrasted the average change in outcomes over time for the treatment group with the average change over time for the comparison group. This approach identified program impacts after separating out any differences between the two groups at baseline, as well as general trends in the study areas (captured by the comparison group)—the latter are changes occurring due to things unrelated to the program being studied, such as the effects of contemporaneous programs. The DID analysis estimated the effects of *Ndaw Wune* on student outcomes by comparing before and after changes in outcomes for the treatment and comparison group students.

Two key features of the DID estimator are particularly attractive for deriving unbiased program impacts. First, using pre- and post-intervention measures enables us to "difference" out unmeasured, fixed (i.e., time-invariant) individual and school characteristics that may affect outcomes, such as gender or school location. The approach also enables us to benchmark the change in the indicator against its value in the absence of the intervention. Second, using the change in a comparison group as a counterfactual enables us to account for general trends in the value of the outcome. The key assumption underpinning the DID approach is that there is no systemic, unobserved, time-varying difference between the treatment and comparison groups. The key identifying assumption is that the trends in the outcomes in the treatment and comparison schools were similar before the start of the program.[15] We believe this assumption is credible because the treatment and comparison schools had mostly similar characteristics before the start of the program (see **Exhibits 6** and **7** above).

Analysis

We assessed the impact of the *Ndaw Wune* program using a DID regression framework while controlling for student characteristics. The sample was analyzed using the following equation:

[14] As a note, given the very specific nature of student selection for this program, this study is only representative of the population sampled for *Ndaw Wune* (mostly CE1 students).

[15] Given the short time frame of the evaluation (i.e., 1 school year), we were not able to test the parallel trends assumption. That is, the DID approach assumed that no time-varying trends exist between the treatment and comparison groups. To assess the parallel trends assumption, data from a few time periods before the treatment would be needed.

$$Y_{ist} = \alpha + \beta_1 Treatment_s + \beta_2 Post_t + \beta_3(Treatment_s * Post_t) + \delta Z_{ist} + \varepsilon_{ist}$$

Where:

- Y_{ist} is the outcome of interest (e.g., reading comprehension) of student *i* in school *s* at time *t*;
- $Treatment_s$ is an indicator variable equal to 0 for schools in the comparison group (no intervention) and equal to 1 for schools in the treatment group (*Ndaw Wune*);
- $Post_t$ is 0 for baseline (pre-intervention) and 1 for endline (post-intervention);
- Z_{ist} includes a set of individual student-level variables that are included as control variables (grade, indicator variables for female, student receiving outside tutoring, and the language the survey was administered in); and
- ε_{ist} is the error term.

The DID design provides us with the intent-to-treat (ITT) effect of the *Ndaw Wune* program (β_3), in other words, the average treatment effect for those students or schools assigned to a treatment condition regardless of take-up of treatment. We analyzed the data in full and by disaggregating it by gender, region, and language of instruction. We also provided robustness checks by controlling for additional school- and student-level covariates, which were statistically different between treatment and comparison groups at baseline.

Sample Sizes and Data Collection Timing

Our team collected baseline quantitative data from treatment and comparison students from November 27–December 1, 2024, right before the program implementation began in December 2024. Follow-up data to measure one-school-year impacts[16] were collected from May 8–June 10, 2025. The original plan was to follow the same students from baseline to endline in a longitudinal design. At baseline, across 70 schools with the target of 1,120 students, we sampled 1,109 students (561 comparison and 548 treatment). Students in each treatment and comparison school were determined based on Step 2 noted earlier (i.e., those students who fell below the cutoff in the initial testing). However, at endline, we were only able to survey 77% of the original baseline sample. At endline, for the 23% of students we were unable to track from baseline, replacement students were selected to keep the endline sample sizes similar to baseline. Replacement students were selected from the original baseline selection tests to ensure that only students below the cutoff at baseline were selected as replacements.[17] In the final analysis, we ran robustness checks with both, the matched sample (students present at both baseline and endline) and the full sample (students at both baseline and endline and replacement students). **Exhibit 8** details the sample sizes for baseline and endline, including the missing and replacement samples.

The overall attrition rate in the sample was 23%, with the treatment group attrition being 25% and comparison group attrition rate at 20%, leading to a 5% differential attrition rate.

[16] Where 'one school year' is equivalent to about 6-7 months (or 22-28 weeks) in this context.

[17] In case students who were below the cutoff at baseline were not available as replacement students on the day of data collection, the selection test was redone, and those below the cutoff were selected as replacements. Since it was the end of the school year, it is possible that those selected as replacements based on the selection test done at endline would be different from those at baseline, since a whole school year of learning would have happened. We examine differential attrition and also conduct robustness analysis with only the matched sample (students at both baseline and endline) and the full samples (original students plus replacements students).

This combination of overall and differential attrition does not exceed the What Works Clearinghouse cautious and optimistic boundaries for high differential attrition.[18] In Section 6, we present baseline equivalence for the full sample, the matched sample, and the missing vs matched students. We also conduct robustness analysis with only the matched sample (students at both baseline and endline) and the full samples (original students plus replacements students). In the full sample analysis, the samples are treated as repeated cross-sections (i.e., no imputation of baseline values is done for the replacement sample).

Exhibit 8. Student Sample Sizes

Group	Target sample	Actual baseline sample	Matched endline sample	Replacement students	Full endline sample
Comparison (35 schools)	560	561	446	111	557
Treatment (35 schools)	560	548	412	149	561
Total	**1,120**	**1,109**	**858**	**260**	**1,118**

Outcome Measures

We collected quantitative data through (a) student surveys and (b) student literacy and numeracy assessments.

Student survey. AIR administered a survey to students and gathered information on key topics, including background information and home language and literacy environments. We drew on various theoretical frameworks that describe how school climate is associated with student development, such as ecological systems theory (Bronfenbrenner, 2000), which posits that student learning development is influenced by classroom, school, and community contexts. Enumerators verbally administered this survey using child-friendly language in the national language that the children use and understand well.

Literacy and numeracy assessments. To assess foundational literacy in the national languages that were being used for instruction, we used a foundational literacy instrument based on our foundational learning theoretical framework as well as on psychometrically validated and locally adapted tools that AIR and other partners have previously utilized in Senegal. These tools assessed children's phonological skills, language comprehension (including receptive and expressive vocabulary and listening comprehension), symbol/letter sound knowledge (sound and name, as appropriate), decoding of real and non-real words, oral reading fluency, and reading comprehension skills. To assess foundational numeracy, we used an adapted version of the Early Grade Mathematics Assessment (EGMA) to assess students' oral counting, number identification, quantity discrimination, missing numbers, and addition and subtraction skills. **Exhibit 9** outlines the constructs covered in the literacy and numeracy assessments, and a translation of the full assessments is presented in **Appendix F**. All tests were conducted in the language of instruction, which was the Senegalese national language selected as the official language of the school.

[18] The cautious and optimistic boundaries for differential attrition for a 23% overall attrition rate are 5.1% and 9.5% respectively (*What Works Clearinghouse ™ Standards Handbook Version 4.0*, n.d.).

Exhibit 9. Literacy and Numeracy Assessment Constructs

	Construct	Subtask (if different from construct)	Description
Literacy	**Listening comprehension**		The child listens to a short story, read to them twice, and responds to corresponding comprehension questions.
	Language dominance (expressive vocabulary)	**Semantic fluency**	The child names as many words in all the languages they claim to know (even a little bit of) as they can within one minute, using an image of a familiar scene as a prompt, then repeats the activity in other languages they know.
	Expressive phonological awareness		The child listens to familiar words read aloud to them, then identifies the beginning sound (phoneme) of each word.
	Letter sound knowledge		The child sees a list of letters (including some digraphs, diphthongs, and letter blends) and produces the sound that is represented by the letter/letter cluster.
	Grapheme-phoneme correspondence Knowledge (decoding)	**Decoding familiar words**	The child sounds out a list of words commonly used. in the targeted language.
		Decoding invented words	The child sounds out a list of pseudo words that follow the orthographic patterns of the targeted language.[19]
	Oral reading fluency		The child reads a short story aloud, and the enumerator notes the correct number of words sounded out with speed and accuracy in 60 seconds.
	Reading comprehension	**Reading comprehension**	The child responds to comprehension questions related to the short story from the oral reading fluency subtask (only for the sections they read up to).
Numeracy	**Number identification**		The child orally identifies written numbers in standard form, progressing from one-digit to three-digit numbers.
	Counting		The student looks at images with repeated objects and counts how many they see.
	Number discrimination		The child is shown a pair of numbers and identifies which of the numbers is bigger, progressing from one-digit to three-digit numbers.
	Number sense	Missing number task	The child is shown a list of four numbers in a pattern, with one number missing. The child identifies the missing number.
	Addition		The child responds to addition problems with progressing difficulty.
	Subtraction		The child responds to subtraction problems with progressing difficulty.

[19] This subtask distinguishes whether a child can decode, rather than knows, familiar words by sight.

Data Collection

Before baseline data collection, AIR received IRB approval from AIR's internal IRB. AIR also explored local IRBs but found that it was not required because the study was not collecting medical data. The data collection team of 16 enumerators and four supervisors attended an in-person training in Dakar led by Dalberg and joined remotely by AIR. Stakeholders from ARED also attended the training and provided valuable input. Enumerators were recruited specifically based on the language of instruction used in each region. Prior to recruitment, Dalberg had a clear understanding of which language would be used in each school and commune. This allowed them to select enumerators who were not only speakers of these languages but also fluent in the local dialects used within the selected communes. This dialect-level fluency was essential to ensure clear communication and full comprehension between enumerators and students to ensure a common understanding of each sub-task being tested between students and enumerators. Additionally, all recruited enumerators were bilingual or multilingual, each spoke at least two languages in addition to French, including Wolof and either Serer or Pulaar. This linguistic versatility was a key factor in maintaining high-quality data collection across diverse linguistic contexts.

The training, which lasted from November 20–22, 2024, included one day of piloting the tools at a non-program school in Dakar with enumerators practicing the tools at least once. Baseline data collection lasted from November 27–December 21. Dalberg supervisors called school principals in advance to schedule the visits so as to minimize interruptions. Data was recorded on tablets via Tangerine software. Endline data collection followed the same procedures as baseline and lasted from May 8–June 10. At endline 13 of the 16 enumerators returned with three replacements. A two-day refresher training took place on May 5 and 6, 2025.

After receiving the cleaned baseline and endline datasets from Dalberg, we performed quality checks to ensure the accuracy of the data. Specifically, we used Stata to look for outliers, nonsensical responses, duplicated observations, and more. Any concerns (e.g., potential duplicates, students coded as taking the survey in French, and inconsistent stop rules) were sent back to the data collector and were resolved with the help of their data collection managers. After we were satisfied with the dataset, we calculated the outcome variables for each of the literacy and numeracy subtests and standardized the results. In addition to quality checks, AIR ran an inter-rater reliability (IRR) analysis to check for consistency between enumerators which yielded positive results (see **Appendix E**). To do so, several surveys (15% at baseline and 20% at endline) were conducted with a speaker and an observer where the speaker administered the survey as usual and the observer watched and recorded their own responses in their tablet without contributing to the questioning. Their responses were compared using an Analysis of Variance (ANOVA) regression. The analysis found a high level of consistency between enumerators for every outcome checked.

ARED flagged certain items in the subtests as too difficult, but to maintain comparability between baseline and endline we decided to not change the test. Instead, we ran the psychometric analysis as a robustness check. The analysis did not reveal a strong relationship between items flagged by ARED and by psychometric analysis. Specifically, to assess the quality of the test items, we conducted a series of psychometric analyses using both classical test theory (CTT)[20] and Rasch modeling.[21] These methods help us identify items that may not perform well and should be excluded from scoring. We conducted Rasch modeling in two ways: one that included all items across languages (concurrent calibration) and another that examined items separately by language. Each approach offers complementary insights into item difficulty and performance. We conducted Rasch modeling with concurrent calibration using the TAM package in R to estimate item difficulty across all languages within each domain (e.g., listening comprehension). This approach allows us to compare item difficulty across languages within the same domain—for example, an item with a difficulty score of −1 in Pulaar is easier than an item with a score of 0 in Serer. However, comparisons across different domains (e.g., listening vs. reading) are not valid. We also conducted Rasch analyses by language and domain using the eRm package in R. While this approach does not allow for direct comparisons of item difficulty across languages or domains, it does provide valuable fit statistics that help us evaluate whether items behave as expected.[22] The assessment demonstrated sound psychometric performance across languages, supporting its suitability for the impact evaluation. Convergent findings from classical test theory and Rasch analysis indicate that the modules are well targeted. Item difficulties cluster in the desired mid-range, Rasch difficulties are moderate, and Infit mean squared values fall within recommended bounds of 0.7 to 1.3. Robust sample sizes in these domains further enhance the stability of estimates, while the picture vocabulary and comprehension skills show substantial *N*s with only scattered misfit. Together, these features indicate internal-structure validity and that score reliability is satisfactory across Pulaar, Serer, and Wolof.

Limitations of the Impact Evaluation

The impact evaluation faced a few limitations. First, there was limited school-level data available for matching treatment and comparison schools. Ideally, we would have been able to match based on observable characteristics such as student and teacher attendance rates, dropout rates, existence of school meals, and other school-level resources, and, importantly, the language of instruction at the school. However, much of this data was not available prior to data collection when matching was conducted. Thus, while the school-level matched sample is balanced on the limited variables that were available, the matches could potentially be different for many other observable and unobservable characteristics. Second, we had limited statistical power to detect heterogeneous effects. While our sample size was powered for overall effects, our evaluation design includes a limited number of schools, constraining our ability to detect heterogeneities in the impact estimates with sufficient precision. Additionally, a key assumption when using the difference-in-differences methodology is that treatment and comparison students were on identical tracks before

[20] Under classical test theory, item difficulty is defined as the proportion of respondents who answered the item correctly. Higher values indicate easier items, while lower values suggest greater difficulty. Ideally, items should fall within a difficulty range of 0.2 to 0.8. Items with values below 0.1 or above 0.9 are flagged as potentially too difficult or too easy, respectively, and are reviewed for exclusion.

[21] Rasch modeling is a statistical approach used to estimate item difficulty and person ability on a common scale, where item difficulty reflects the point at which a person has a 50% chance of answering correctly. Difficulty scores are expressed as z-scores, where negative values indicate easier items, zero represents average difficulty, and positive values indicate harder items. Ideally, scores should fall between −2 and +2.

[22] Specifically, we examined two types of fit statistics: *infit* and *outfit*. Infit is sensitive to unexpected response patterns on items that match a person's ability level, while outfit is more sensitive to unusual responses on items that are very easy or very hard for a person. These statistics help us determine whether items are functioning appropriately, for example, whether more difficult items are consistently answered correctly by individuals with higher ability.

implementation began (i.e., without *Ndaw Wune*, the treatment and comparison students would have had the same outcomes). This is usually tested by collecting data from multiple points before implementation begins. It is common (as in this case) that previous data collection is not possible, and the matching strategy and baseline equivalence tests help to address this limitation. Finally, we did not have access to language of instruction information about schools before sampling was done. In the end, we had data for 12 Serer treatment schools and only one Serer comparison school. This was not a large enough sample to make reliable comparisons between the treatment and comparison groups. The sample included 12 Wolof treatment and 23 Wolof comparison schools and 11 Pulaar treatment and comparison schools.

5.2. Analysis of Monitoring Data

To supplement the evaluation, ARED shared a subset of data from their continuous monitoring exercises. These data included (a) frequency counts of student absences from December 2024 through March 2025 and (b) brief commentary from 54 teachers (across 10 supervisors in the four target regions) related to the program from December 2024 through April 2025 (with 552 overall observations). AIR researchers used the teacher commentary to provide additional insights related to the program operation. However, the student absence data included several inconsistencies,[23] which precluded us from drawing any reliable insights.

Analysis Method

Our team followed a systematic process to qualitatively analyze the teacher commentary. Using NVivo qualitative research software, we first uploaded the cumulative commentary and classified the data by (a) supervisor, (b) region, and (c) school. We then inductively coded the commentary according to emergent themes.

Limitations of the Monitoring Data

Our use of ARED's monitoring data in this evaluation is subject to the following limitations:

Partial coverage of the implementation timeframe. We did not receive any teacher commentary after April 2025, so were not able to include data from the last month of implementation.

Brevity of teacher commentary. The comments we received from teachers included brief statements related to program implementation, such as "Problem understanding certain exercises" or "Appreciation from parents." While this data provides helpful information about the successes and challenges teachers and students faced, it does not allow us to elaborate extensively or provide details such as the exercises that teachers and students struggled to understand.

5.3. Cost Analysis

To respond to Evaluation Questions 2–4, AIR conducted a retrospective programmatic cost analysis of the *Ndaw Wune* remedial education program. That is, we collected costs from the program perspective, including all costs directly incurred by the program implementer.

[23] Inconsistencies included missing reports from *Ndaw Wune* instructional days and missing reports from tutors and schools within the included days.

Our team produced cost estimates for the 2024–2025 school year version of the program: *Ndaw Wune Année Scolaire.*

Time Horizon. We based our cost analysis on programming that occurred during the 2024–2025 school year. We used costs related to the regions and numbers of students and schools that the program served during that timeframe. To better align with what would be considered "normal" programming, we made a few costing adjustments. For example, during the 2024–2025 school year, ARED adjusted their regular evaluation procedures (i.e., they conducted only a midline evaluation using an adjusted sample size) to align and avoid redundancy with AIR's impact evaluation. Therefore, we based the programmatic evaluation costs on procedures likely to occur during a normal school year (i.e., two rounds of data collection), using the school locations and prices used for the program in the 2024–2025 school year. Also, we updated all prices to the current G50 protocol to reflect typical prices going forward.[24]

Approach. AIR adopted a hybrid approach to the cost analysis, incorporating both ingredients-based and activity-based costing methods.

The ingredients approach is a resource-intensive best practice methodology (Danks & Kendziora, 2021; Levin et al., 2018), which involves a systematic process to identify and quantify every program input (such as number of person-hours by role and activity, type and number of materials used, etc.) and then assign a unit price to each input to calculate the input cost. The ingredients-based approach allows for a high level of precision in the decomposition of costs to constituent activities, but this method is highly data-intensive and requires significant input from all program staff and participants to collect in-depth information about the specific quantities and types of resources used. To optimize cost efficiency and minimize data collection burden among staff and participants, we limited our data collection for the cost study to programmatic and financial documents that ARED provided to us and interviews with key ARED staff. Using the data provided, we followed the ingredients-based approach for resources for which we were able to obtain detailed unit price and quantity data, such as the number of student booklets and travel reimbursement costs for teachers during training.

For program inputs lacking ingredient-level data, we supplemented with an activity-based approach. The activity-based approach involves the identification of necessary program inputs for each program activity and the subsequent allocation of value, using program budgets, expenditure documents, activity records, and key informant interviews (Brander et al., 2024). That is, instead of gathering data about the number of hours each ARED staff spent on each activity within *Ndaw Wune* and then multiplying that quantity by the labor price per hour to determine personnel costs for each program component, the activity-based approach uses program activity descriptions, annual salary data, and interviews with ARED organizational staff to proportionally allocate an estimated personnel cost to each program activity. Specifically, activity-based costing was used to allocate ARED personnel costs across the four program components and to estimate travel costs when itineraries

[24] The G50 refers to the *Groupe élargi de concertation et de coordination des partenaires au développement du Sénégal*, a nationally recognized group that standardizes travel reimbursement and stipend rates for government and project personnel.

were provided without accompanying expenditure data. This hybrid approach is an efficient methodology to produce rigorous and usable cost data that can inform program scale-up.

Data Sources

To obtain cost information, our team (a) received program and organizational budget and expenditure data from ARED and (b) conducted interviews, held extensive discussions, and corresponded via email with relevant program staff. Our communications with ARED included the following topics:

- The program components and associated tasks and activities,
- The staff involved in the program, their rates, and the time spent on the program,
- The facilities, equipment, and materials used for the program and their costs, and
- Resources used for program administration.

The data we received from ARED included the following:

- Salary information for ARED staff and the percentage of their time they devoted to *Ndaw Wune,*
- Printing and delivery costs of materials,
- Numbers of schools, students, teachers, and school locations for *Ndaw Wune Année Scolaire,*
- A list of planned activities for *Ndaw Wune Année Scolaire,*
- Terms of Reference (TOR), itineraries, and/or budgets for teacher trainings, pedagogical days, and evaluations, and
- The government protocol for travel reimbursements (the G50 protocol).

Using informants' estimates and expenditure data from ARED, we captured retrospective data on the ingredients and quantities of those ingredients used to implement *Ndaw Wune Année Scolaire* in 2024–2025. When ingredients-level cost data were missing, we made assumptions to best estimate program costs. For example, we received the TOR and itinerary for one round of pedagogical days (but no budget document), and we used the travel budget for an evaluation round, supplemented with conversations with ARED, to estimate the travel costs associated with the pedagogical days. We documented all assumptions we made (presented in **Appendix C**), which we shared with ARED program and finance staff to verify and/or correct via email and through two virtual meetings.

Analysis

AIR's cost analysis team recorded cost data in a database called a resource cost model (RCM). In the RCM, we listed costs for each of the following program activities:

- **Initial teacher training**: The 4-day teacher training that occurs at the beginning of each program cycle, preceded by a 1-day "harmonization day" for the trainers.
- **Direct instruction**: The weekly classroom instruction for *Ndaw Wune*, which includes costs of instructional materials and teacher pay, among others.

- **Pedagogical days**: The half day refresher training, preceded by classroom observations, which occur twice during the school year.
- **Management, planning, and monitoring & evaluation**: All other costs associated with program implementation, such as time used for office work related to furnishing TLMs, preparing for training, and engaging in ongoing monitoring and evaluation. This category also includes costs for two rounds of student assessments that ARED conducts when resources allow.

We separated costs into five categories, keeping each resource as a separate line item within each category:

- **Personnel:** All staff time dedicated to the program, including ARED program staff, consultants to facilitate trainings, and stipends for teachers, inspectors, and supervisors.
- **Materials and equipment:** Materials (e.g., teacher and student books) and equipment (e.g., laptops and projectors) used in the operation and implementation of the program.
- **Travel:** All expenses related to transportation and accommodation, such as mileage reimbursements, driver and fuel fees, lodging, and meals/per-diems.
- **Facilities:** Any costs to the program for the use of physical space. Note that use of classrooms for direct instruction, teacher training, and pedagogical days incurred no cost to the program and were thus excluded. We do account for the costs to rent space for the evaluator training. Costs of office space are included in the overhead category, described below.
- **Overhead:** Costs such as office space and supplies and other general organizational operations expenses. ARED budgets for these expenses by allocating 10% of project costs to overhead, which we mirrored in the cost analysis.

We then attached each ingredient to its price and quantity. The RCM then calculates the costs of each input, summarizing across resource categories and program components, and presents results in both FCFA and USD. Using these aggregate cost estimates, we divided by 3,000 to estimate per-student costs.

Although we originally planned to separately calculate costs for annual operations, set-up (to provide trainings and materials in new schools in the same region), and scale-up (to translate materials and secure new partnerships), we learned through an interview with the *Ndaw Wune* program manager that, as the program is still in an "action research" phase, materials have been revised each year to date. Therefore, we included these materials revision costs in our cost estimate, even though the materials will be stabilized after revisions in 2026 and this will not be an ongoing cost.

Limitations of Cost Analysis Data

As with any study, our cost analysis is subject to some limitations and risks, which we note as follows:

Hybrid approach. While our hybrid approach of combining ingredients-based and activity-based costing is both appropriate and efficient for this analysis, there is a trade-off in terms of precision. For program activities for which we did not have detailed expenditure data, we estimated costs based on activity descriptions and TORs. Program documents contained limited detail on the disaggregation of personnel costs and transportation expenditures, leading us to make several consequential assumptions for these cost categories. We shared our preliminary assumptions with ARED via email and a virtual meeting, asked for clarifications and additional data, and modified our assumptions based on ARED's feedback. We then held an additional meeting with ARED in which they provided the precise personnel costs that should be allocated to *Année Scolaire.* All assumptions are thoroughly documented in **Appendix C**.

Recall bias. Our cost data is subject to recall bias of program implementers. Triangulation between the documented cost data and the interview data helped to mitigate the drawbacks of recall bias. In cases of discrepancies between the documented data and interview data, we deferred to the documented data.

Fluctuating exchange rate. Following best practices, we conducted the cost analysis in local currency (FCFA), and we used an exchange rate of 580 FCFA per 1 USD because that is the rate embedded in the budget spreadsheets received from ARED. However, the exchange rate fluctuated widely throughout the period of program implementation—from around 642 FCFA per 1 USD in February 2025 to the current rate of about 563 FCFA per 1 USD. Thus, USD cost estimates should be interpreted with caution.

5.4. Cost-Effectiveness Analysis

Using results obtained from the cost analysis and impact evaluation, we assessed the observed impacts of *Ndaw Wune Année Scolaire* relative to the costs of implementing the program. In a cost-effectiveness analysis, researchers typically compare the costs of the program to "business as usual"—the opportunity costs or comparison conditions that students would otherwise receive. The "business as usual" scenario in this case is no remediation program. Thus, the "business as usual" costs for comparison are $0.

Because the impact evaluation assessed outcomes across multiple literacy and numeracy competencies, we conducted a type of cost-effectiveness analysis that is better suited to multi-objective programs, often called a "cost-consequence analysis" or "impact inventory." Such an approach is more relevant for the *Ndaw Wune* cost analysis because the costs to achieve the various literacy impacts and numeracy impacts are not separable; for example, we cannot partition costs to say "x% of the program costs lead to literacy impacts and y% of the program costs lead to numeracy impacts." Therefore, we do not present ratios of costs to impact as in a traditional cost-effectiveness analysis but instead display program costs alongside the inventory of program impacts. To facilitate comparability with the costs and impacts of other remedial education programs, we also present a linear transformation of per student impacts achieved per $100 investment.

6. Findings

6.1. Impact Evaluation

Student-Level Baseline Equivalence

We presented school-level balance in **Exhibits 4** and **5**; in this section, we start by examining student-level baseline equivalence for both the full sample and only the matched sample as well. As noted in Section 4.1, matching was conducted at the school level, resulting in a relatively balanced school-level sample. Students in treatment and comparison groups were then selected based on the screening tests for literacy and numeracy, with tests being administered in the same way in both treatment and matched comparison schools. **Exhibit 10** presents the baseline differences in observable characteristics between students in treatment and comparison schools based on the full baseline data. There were some notable differences between the treatment and comparison students at baseline, language being a critical one. While the treatment group was almost evenly distributed between Wolof, Serer, and Pulaar, the comparison group was made up of just 3% Serer speakers (from one school), 31% Pulaar speakers, and 66% Wolof speakers. Additionally, 88% of treatment group students said they were receiving outside tutoring, while 76% of the comparison group said they received the same ($p < .01$).[25] The literacy test at baseline revealed one difference between the groups in letter sound knowledge—the comparison group outperformed the treatment group by 4.6 letters (out of 100; $p < .10$). There were several differences for the numeracy sub-skills. The treatment group scored higher than the comparison group in counting, digit discrimination, and addition; however, the magnitude of the difference is quite small for each—under one correct answer.

Given the differences in observable characteristics between treatment and comparison students (despite selecting students using the same test), we present robustness checks, including all baseline covariates that are statistically significant as controls in the main estimation (**Exhibit D6, Appendix D)**.

Exhibit 10. Student-Level Baseline Characteristics

Outcome	Comparison group	Treatment group	Difference
Girl	50%	50%	0%
Age	9.51 years	9.64 years	0.14 years
Grade	95%	96%	1%
Attended a preschool	58%	59%	2%
LOI: Wolof	66%	33%	−33%***
LOI: Serer	3%	35%	32%***
LOI: Pulaar	31%	32%	1%
Eaten food today	83%	82%	−1%

[25] It is possible that students in treatment schools misunderstood and considered *Ndaw Wune* as outside tutoring.

Outcome	Comparison group	Treatment group	Difference
Received tutoring outside of school	76%	88%	12%***
During the last full week of school, how many days did you miss?	1.56 days	1.38 days	−0.18 days**
Listening comprehension	3.14	3.4	0.27
Phonological awareness	7.59	8.46	0.87
Letter sound knowledge	75.31	70.75	−4.55*
Decoding invented words	6.64	5.99	−0.65
Decoding familiar words	6.11	6.74	0.64
Oral reading fluency	5.24	5.57	0.33
Reading comprehension	0.54	0.68	0.14
Number ID	18.07	18.23	0.16
Counting	5.11	5.33	0.22**
Digit discrimination	9.74	10.39	0.66**
Number sense	4.43	4.48	0.05
Addition	6.21	7.01	0.80*
Subtraction	4.76	5.12	0.35

Note. LOI = language of instruction. The sample includes all students surveyed at baseline.
*$p < .10$. **$p < .05$. ***$p < .01$.

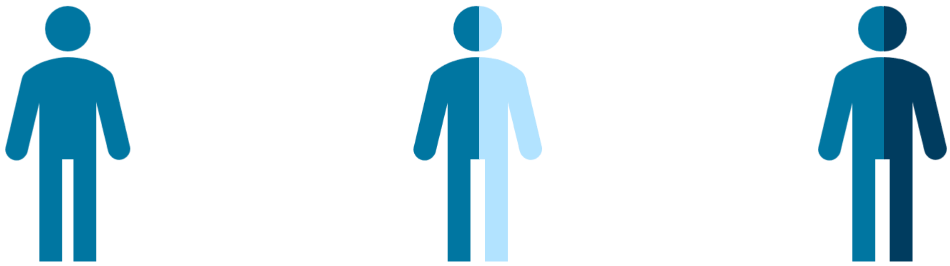
Additionally, as noted earlier in Section 5.1, there was 23% overall attrition and 5% differential attrition. To make sure there was no systematic reason for the missing students at endline, we analyzed the baseline equivalency of the matched versus missing students (see **Exhibit D2, Appendix D**). There were only a few differences in the treatment sample—more of the missing sample were Wolof speakers compared to the matched sample (43% vs. 30%, $p < .10$), and more of the matched sample were Pulaar speakers (35% vs. 23%, $p < .05$). Missing students were more likely to be in Kaolack (37% vs. 20%, $p < .05$) and less likely to be in Saint Louis (10% vs. 24%, $p < .05$). There was no difference in Matam. Missing students (at baseline) also self-reported missing 0.3 more days of school the previous week compared to matched students ($p < .05$). None of the outcome variables related to literacy or numeracy differed between missing and matched treatment students.

Among comparison students, missing students were more likely to be girls (65%) compared to matched students (47%, $p < .01$). Additionally, missing students were more likely to be in Kaolack (44% vs. 29%, $p < .01$) and less likely to be in Saint Louis (30% vs. 22%, $p < .05$), with no difference in Matam. For the foundational literacy scores, the only difference was that matched students were, on average, able to read 1.5 more familiar words (out of 50) than the missing students ($p < .10$). Given some of the differences in observable characteristics between the matched and missing samples, as a robustness check, we ran the impact regressions with the full and matched samples and found similar results. Additionally, we also present the baseline equivalence with only the matched sample in **Exhibit D3, Appendix D** and with the full sample without the Serer sample in **Exhibit D4, Appendix D**.

Descriptive Analysis of Semantic Fluency

As part of both baseline and endline data collection, we empirically determined how well students spoke all the languages they claimed they could speak. We utilized semantic fluency assessments, which measure expressive oral vocabulary to establish which language(s) students had dominance in case any child was being taught in a national language that was not their own home language. Using this data, children were grouped into three categories: monolingual, emerging bilingual, and bilingual. Monolingual students were those who claimed to know only one language, while emerging bilingual students were those who claimed to know two languages and scored more than five words apart in the number of expressive vocabulary items they were able to produce in each language within the 60-second window. Bilingual students were those who were able to produce within five words of each language they claimed to know. This was an important variable to ensure students were actually tested and taught in languages they do understand and speak dominantly (as a monolingual or as a bilingual child). Definitions for classification of children into different categories of language abilities are described in **Exhibit 11**.

Exhibit 11. Language Classifications



Monolingual	Emerging bilingual	Bilingual
Child claims to know only one language.	Child claims to know more than one language but scored more than a five-point difference between the languages.	Child claims to know more than one language and scored within five points in both languages.

According to the semantic fluency test, most children (93% at baseline and 94% at endline) were monolingual in a national language. For instance, at baseline 87% and 91% of students were monolingual in Wolof at baseline and endline, respectively. A small fraction of students were bilingual (7% of the students sampled at baseline and 5% at endline), and less than 1% were emerging bilingual. Among bilingual students, the proportion of urban students was higher. The most known language was Wolof, followed by Pulaar and Serer (**Exhibit 12**). Specifically, **Exhibit 12** shows the distribution of languages that students know in the sample. For instance, 51% of the students have Wolof as one of their dominant languages at baseline and at endline. Most of these students are monolingual – 87% of the Wolof-speaking students (51% of the sample at baseline and endline) are monolingual at baseline and 91% of them are monolingual at endline. Similarly, 86% of the Pulaar-speaking students (34% of the sample at baseline and endline) are monolingual at baseline and 87% of them are monolingual at endline.

Exhibit 12. Language Dominance in Students

Language	% dominant in the language at baseline	*N*	% dominant in the language at endline	*N*
Wolof	50.77	1,109	51.16	1,118
Pulaar	33.63	1,109	34.26	1,118
Serer	18.94	1,109	18.52	1,118
French	1.89	1,109	0.36	1,118
Sonike	1.44	1,109	0.36	1,118
Bambara	0.27	1,109	0	1,118
Bassari	0.09	1,109	0.09	1,118
Diola	0.09	1,109	0	1,118

Almost all students were dominant in the national language that was the language of instruction for the school. The level of mismatch between a student's dominant language and the language of instruction was 3% at endline and 2% at baseline. The level of mismatch is low across the three languages of instruction – Wolof, Serer, and Pulaar (**Exhibit 13**).

Exhibit 13. Match between Students' Dominant Language and the Language of Instruction

Mismatch %	Endline	N	Baseline	N
Wolof	2%	558	4%	552
Serer	0%	206	2%	208
Pulaar	1%	354	4%	349

Overall Treatment Effects

Exhibit 14 shows the impact of *Ndaw Wune Année Scolaire* on the literacy and numeracy skills noted in Exhibit 9 above. These are presented as standardized effect sizes[26] in column six. Each cell in column six is a separate regression estimate. Any row that is highlighted in green and with asterix is a positive and significant finding. All regressions control for the following: grade, gender, whether the student was receiving outside tutoring, and the language the survey was administered in.[27] Results are presented for the full sample (students matched between baseline and endline and replacement students).

Overall, the data show **that *Ndaw Wune Année Scolaire* had a positive impact on foundational literacy sub-skill scores and reading comprehension scores.** The largest impact was on familiar word decoding,[28] followed closely by letter sound knowledge. In terms of **familiar word decoding**, students could, on average, decode six of 50 familiar words at baseline, while treatment students could decode seven. At endline, treatment

[26] Calculated by subtracting the endline mean from the baseline mean and dividing by the standard deviation at baseline.
[27] We do not include days absent as a control since this was self-reported and therefore not very reliable. We do include it as a control in the robustness checks section.
[28] "Familiar words" refers to actual words that would appear in the students' curriculum which they are expected to be able to decode.

students improved by 12 percentage points to 13 words, while comparison students improved by six percentage points to nine words, equating to a treatment effect of 0.47 *SD* ($p < .01$). For **letter sound knowledge** at baseline, comparison students could identify the sounds of 75 out of 100 letters compared to 71 for treatment students. By endline, comparison and treatment students increased their letter sound knowledge scores to 78 and 84 out of 100, respectively, equating to a 0.45 *SD* treatment effect ($p < .01$). At baseline, comparison students could decode an average of seven of 50 **invented words** compared to six for treatment students. By endline, comparison students increased their scores to 10 while treatment students increased to 12 words, which translated to a 0.32 *SD* treatment effect ($p < .01$). These three decoding skills combine for a decoding treatment effect of 0.42 *SD* ($p < .01$).

There was also a small but positive and significant impact of *Ndaw Wune* on students' **oral reading fluency** scores**.** Comparison students at baseline could read five of 50 words, compared to six for treatment students. At endline, comparison students read 10 words, while treatment students increased their scores to an average of 13 words. The stronger increase in oral reading fluency for treatment students translated to a treatment effect of 0.32 *SD* ($p < .01$). Finally, treatment students' **reading comprehension** also increased. At baseline, comparison students averaged 0.5 correct answers out of six compared to 0.7 for treatment students. At endline, both groups raised their scores to 1.0 and 1.4, respectively, leading to a treatment effect of 0.14 *SD* ($p < .10$).

Two of the five numeracy subtests showed significant impacts of *Ndaw Wune*. The first, **number sense**, asked students to fill in the missing number from a series of numbers, for example: 1, __ , 3, 4. Here, treatment students improved from 4.5 out of 12 at baseline to 5.8 at endline while comparison students from 4.4 to 5.3, which equated to a 0.19 *SD* treatment effect ($p <.05$). The second positive impact was for the **subtraction** subtest, which asks progressively more difficult subtraction questions. Treatment students improved from an average of 5.1 out of 12 to 7.3 at endline while comparison students increased from 4.8 at baseline to 6.2 at endline, for a treatment effect of 0.2 *SD* ($p < .05$).

We also looked at effects for the matched students in treatment schools based on the *Ndaw Wune* numeracy groups in which students were placed at baseline (i.e., students who could do fewer than five number comparisons [i.e., the numbers group] and students who could do all five number comparisons [i.e., the operations group]). Students in both math groups showed statistically significant improved mean scores on every outcome we tested. Students in the numbers group had a larger improvement than the more advanced operations group in four of the six numeracy outcomes measured, including the three lower-order outcomes: number identification, counting, and digit discrimination.

Exhibit 14. Treatment Effects (Full Sample With Controls)

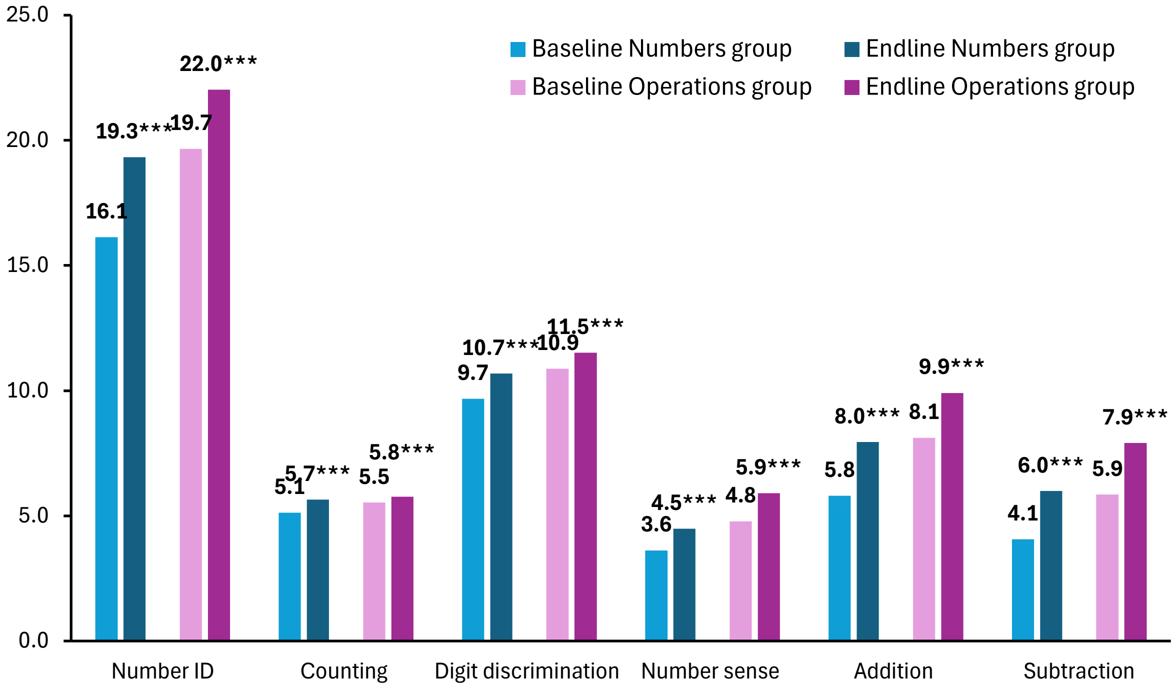
Subtask	Comparison baseline	Comparison endline	Treatment baseline	Treatment endline	Standardized treatment effect	*N*
Listening comprehension (out of 6)	3.14	4.22	3.40	4.41	−0.04 (0.10)	2,227

Subtask	Comparison baseline	Comparison endline	Treatment baseline	Treatment endline	Standardized treatment effect	*N*
Phonological awareness (out of 10)	7.59	7.35	8.46	8.74	0.15 (0.13)	2,227
Decoding construct	n/a	n/a	n/a	n/a	0.42*** (0.08)	2,227
• Letter sound knowledge (out of 100)	75.31	77.69	70.75	83.52	0.45*** (0.10)	2,227
• Invented word decoding (out of 50)	6.64	9.94	5.99	11.85	0.32*** (0.08)	2,227
• Familiar word decoding (out of 50)	6.11	9.00	6.74	12.87	0.47*** (0.10)	2,227
Oral reading fluency (out of 50)	5.24	9.88	5.57	13.12	0.32*** (0.11)	2,227
Reading comprehension (out of 6)	0.54	0.97	0.68	1.35	0.14* (0.08)	2,191
Number ID (out of 24)	18.07	20.34	18.23	21.00	0.11 (0.07)	2,227
Counting (out of 6)	5.11	5.51	5.33	5.69	−0.03 (0.11)	2,227
Digit discrimination (out of 12)	9.74	10.85	10.39	11.22	−0.10 (0.08)	2,227
Number sense (out of 12)	4.43	5.25	4.48	5.84	0.19** (0.08)	2,227
Addition (out of 12)	6.21	8.27	7.01	9.28	0.07 (0.08)	2,227
Subtraction (out of 12)	4.76	6.22	5.12	7.26	0.20** (0.09)	2,227

Note. All treatment effects are standardized. Standard errors are in parentheses. Controls include grade, indicator variables for female, student receiving outside tutoring, and the language in which the survey was administered. The decoding construct is the mean treatment effect of three decoding subtests, including letter sound knowledge, invented word decoding, and familiar word decoding. Oral reading fluency was the only timed test at 60 seconds.
*$p < .10$. **$p < .05$. ***$p < .01$.

Exhibit 15 shows the average score of each numeracy subtest for the numbers and operations groups. For **number identification**, both the numbers and operations group saw significant improvements from 16.1 to 19.3 and 19.7 to 22.0 out of 24, respectively (p < .01). For **counting,** both groups saw significant increases (p < .01), although they were small in magnitude. The numbers group increased from 5.1 to 5.7 out of 6 while the operations group went from 5.5 to 5.8. Again, students in both groups improved slightly but significantly in the **digit discrimination** subtest, with the numbers group improving from 9.7 to 10.7 out of 12 while the operations group score increased from 10.9 to 11.5 (p < .01). For **number sense,** the numbers group improved from 3.6 at baseline to 4.5 at endline out of 12, while the operations group increased their scores from 4.8 to 5.9 (p < .01). Both groups also had significant increases in the **addition** and **subtraction** subtests, with the numbers group increasing from 5.8 to 8.0 and 4.1 to 6.0 out of 12, respectively, while the operations group improved their addition scores from 8.1 to 9.9 and subtraction from 5.9 to 7.9 (p < .01).

Exhibit 1515. Treatment Group Numeracy Outcomes by Math Group



Note. Math groups determined by ARED selection criteria test (see Appendix A). *N* = 141 for the numbers group and 166 for the operations group.
**p* < .10. ***p* < .05. ****p* < 0.01.

Robustness Checks

We ran several robustness checks. First, to account for the nested structure of the data (students within schools) and for within-school correlation, we have included a random effects model. Results from the random effects model are similar to the main results (**Exhibit D5, Appendix D**). Second, we added additional controls for observable school-level variables and student-level variables that were collected at baseline and were significantly different between treatment and comparison schools and students: if a school had a handwashing station and if that handwashing station had soap, and the self-reported average number of school days missed the previous week. Adding these controls made no meaningful difference to the findings (**Exhibit D6, Appendix D**). Next, we ran the same regressions using only the students who were present at baseline and endline (**Exhibit D7,**

Appendix D). The matched sample regressions had very similar results, with the exception of reading comprehension (no longer statistically significant) and number identification (marginally significant). We also ran the regressions on the full sample without the Serer speaking students (**Exhibit D8, Appendix D**). The regressions followed a similar pattern to the others, with all the decoding outcomes being significant and positive. However, in these regressions, listening comprehension had a negative treatment effect and phonological awareness had a positive treatment effect while none of the numeracy treatment effects were significant.

Heterogeneous Treatment Effects

In this section, we examine heterogeneous treatment effects by gender, language, and region. As a note, our sample size is only powered to detect overall effects, and the statistical analysis has limited power to detect heterogeneous effects.

First, we looked at the treatment effects by gender. As shown in **Exhibit 16**, *Ndaw Wune* had positive effects on girls and boys, with boys consistently seeing stronger treatment effects than girls. For the strong literacy findings mentioned earlier (decoding constructs including letter sound knowledge, invented and familiar word decoding, and oral reading fluency), boys and girls both saw significant impacts. At baseline, girls in the treatment group had statistically significantly higher scores than boys in the treatment group for addition, subtraction, and two of the three decoding outcomes. The larger magnitude of the treatment effect for boys could be explained by boys having more room for growth than girls, relative to their baseline values. With respect to numeracy, there was more variation in effects by gender. Girls had a negative treatment effect on the counting exercise (−0.19 *SD*, $p < .10$), while boys had no significant impact in either direction. For the number sense subtest, girls saw a *positive* impact (0.25 *SD*, $p < .05$), while boys did not. On the other hand, boys had positive and significant impacts for both addition (0.2 *SD*, $p < .05$) and subtraction (0.29 *SD*, $p < .05$).

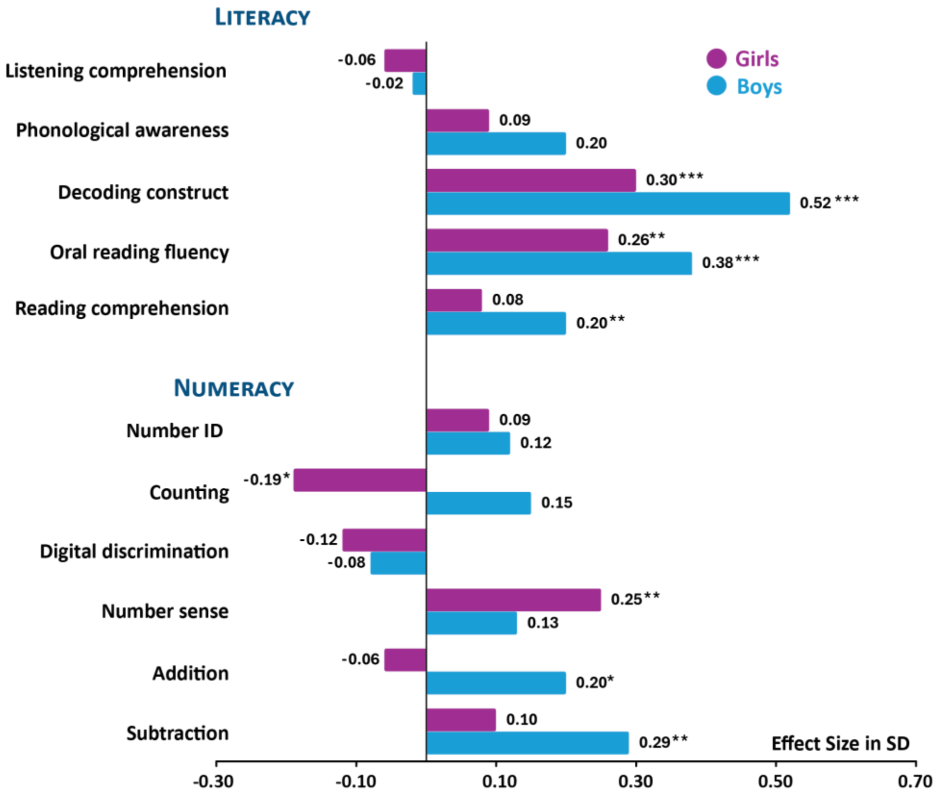
Next, we examine impacts by language of instruction in **Exhibit 17**. As noted in the results presented for semantic fluency, the language of instruction and a student's dominant language are good proxies for each other given the low level of mismatch between them. We show the impact results for Wolof and Pulaar speakers.[29] **The data show that Wolof speakers saw most of the gains from *Ndaw Wune*.** Wolof student impacts were significant and positive for letter sound knowledge (0.27 *SD*, $p < .01$), invented word decoding (0.5 *SD*, $p < .01$), familiar word decoding (0.69 *SD*, $p < .01$), and oral reading fluency (0.53 *SD*, $p < .01$), while Pulaar-speaking students only had significant impacts for letter sound knowledge (0.43 *SD*, $p < .10$). However, Wolof speakers also had a marginally significant negative effect on listening comprehension (−0.3 *SD*, $p < .10$).[30] Wolof speakers did not have any significant

[29] We did not have access to language-of-instruction information about schools before sampling was done. In the end, we had data for 12 Serer treatment schools and only one Serer comparison school. This was not a large enough sample to make reliable comparisons between the treatment and comparison groups. The sample included 12 Wolof treatment and 23 Wolof comparison schools and 11 Pulaar treatment and comparison schools.

[30] At baseline, treatment students in Wolof-speaking schools had high listening comprehension test scores - 4.4 out of 6 compared to 3.4 out of 6 for comparison students. While both groups did improve between baseline and endline, a ceiling effect may have prevented the treatment students from improving at the same or greater rate than the comparison students leading to the negative impact estimate.

impacts in the numeracy exercises, and Pulaar speakers had one subtest, counting, that showed negative effects (−0.37 *SD*, $p < .10$).[31]

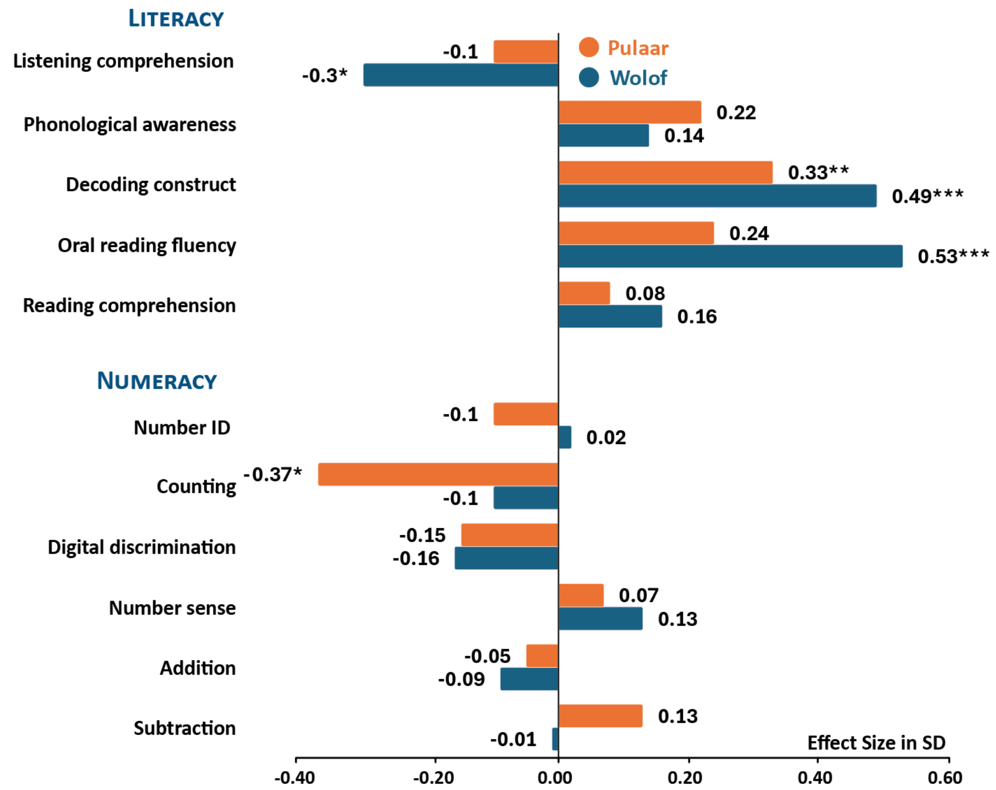
Exhibit 1616. Treatment Effects by Gender



Note. All treatment effects are standardized. Controls include grade, if the student received outside tutoring, and the language in which the survey was administered.
$*p < .10$. $**p < .05$. $***p < .01$.

[31] We checked the matched sample and the results are similar to the full sample. Additionally, there was very low mismatch between the dominant language of instruction and the dominant language spoken by the student (measured through semantic fluency). One possible explanation for the negative impacts on some math subskills could be that, as shown in the teacher commentary data, the numeracy exercises were difficult and sometimes hard to understand for the students. Additionally, since digital discrimination seems to be the sub-skill that seems to have the most consistent negative impact, it would be prudent to revisit simple exercises such as counting and comparing number magnitudes to strengthen and reinforce these concepts.

Exhibit 1717. Treatment Effects by Language of Instruction

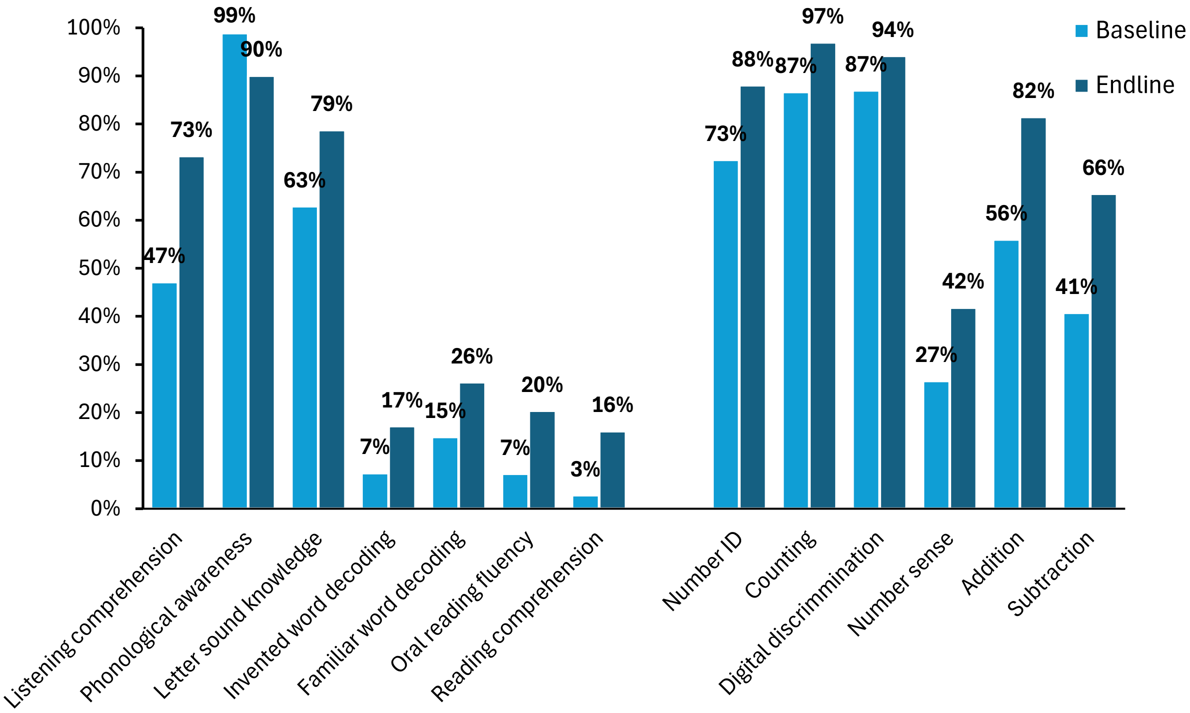


Note. All treatment effects are standardized.
**p* < .10. ***p* < .05. ****p* < .01.

Because we did not have enough comparison data to make reliable comparisons between the treatment and comparison groups for Serer speakers, we examined the pre/post results for only treatment students, comparing treatment student baseline and endline scores on each subtest. **Exhibit 18** shows the baseline and endline mean EGRA and EGMA scores as a percentage. In almost all subtests, students improved positively and significantly from baseline to endline.[32]

[32] For some of the subtests, there were possible ceiling effects that led to smaller, but still significant gains such as counting and digit discrimination, which had strong baseline scores. For one subtest, logical awareness, Serer students averaged a score of 99% at baseline and actually saw a significant decrease at endline, even though their average score of 90% was still quite high.

Exhibit 1818. Baseline and Endline Means for the Treatment Group for Serer-Speaking Students



Note. N = 192 at baseline and 192 at endline (186 for reading comprehension).

Next, we examine impacts by region.[33] **Exhibit 19** shows the impact results by department. The strongest impacts of *Ndaw Wune* came from Kaolack, where students saw positive impacts in four of the six literacy subtests.[34] The treatment effect was particularly strong for familiar word decoding (0.59 *SD*, *p* <.01), followed by oral reading fluency (0.48 *SD*, *p* < .05) and invented word decoding (0.45 *SD*, *p* < .01). Although students in Saint Louis and Matam showed improvement across most subtests from baseline to endline, the treatment effects were noticeably more modest in Saint-Louis and largely absent in Matam.[35] Given that all Pulaar-speaking students in the sample attend schools in these two regions, it is unsurprising that their performance did not match the stronger outcomes observed in the other departments.

[33] We did not have a large enough comparison group sample in Diourbel to make reliable comparisons between the treatment and comparison groups.

[34] Listening comprehension had negative and significant impacts in Kaolack. We checked the matched sample in Kaolack and it had a treatment effect of -0.25 that was significant at the 5% level, which was the same result as the full sample. Since we were not powered for subgroup effects, it could be possible this is an artifact of the fact that this subsample is not powered.

[35] We checked the language mismatch in Matam and there was very low mismatch between the language of instruction and the language spoken by the student. As a note, the analysis was not powered to detect effects at the subgroup level. Thus, an absence of impacts in Matam could be due to the study being underpowered. Finally, in the absence of in-depth qualitative data around implementation fidelity, we are unable to understand the reasons behind low to no impacts in Matam.

Exhibit 191920. Treatment Effects by Region

Outcome	Kaolack		Matam		Saint Louis	
	Treatment effect	*N*	Treatment effect	*N*	Treatment effect	*N*
Listening comprehension (out of 6)	−0.2** (0.09)	630	0.01 (0.15)	511	−0.18 (0.15)	542
Phonological awareness (out of 10)	0.25** (0.09)	630	0.29 (0.2)	511	0.17 (0.16)	542
Decoding construct	0.41*** (0.11)	630	0.20 (0.14)	511	0.40* (0.22)	542
• Letter sound knowledge (out of 100)	0.19*** (0.08)	630	0.43 (0.29)	511	0.32** (0.14)	542
• Invented word decoding (out of 50)	0.45*** (0.14)	630	0.13 (0.12)	511	0.32 (0.25)	542
• Familiar word decoding (out of 50)	0.59*** (0.18)	630	0.06 (0.15)	511	0.57* (0.31)	542
Oral reading fluency (out of 50)	0.48** (0.2)	630	0.03 (0.16)	511	0.31 (0.30)	542
Reading comprehension (out of 6)	0.09 (0.2)	619	0.03 (0.13)	508	0.09 (0.18)	529
Number ID (out of 24)	0.13* (0.07)	630	−0.09 (0.19)	511	−0.01 (0.10)	542
Counting (out of 6)	0.17 (0.11)	630	−0.45*** (0.13)	511	−0.09 (0.16)	542
Digit discrimination (out of 12)	0.08 (0.06)	630	−0.13 (0.14)	511	−0.15 (0.14)	542
Number sense (out of 12)	0.22 (0.14)	630	−0.16 (0.23)	511	0.22 (0.15)	542
Addition (out of 12)	0.13 (0.12)	630	−0.01 (0.12)	511	−0.14 (0.10)	542
Subtraction (out of 12)	0.24 (0.14)	630	0.08 (0.11)	511	0.04 (0.13)	542

Note. All treatment effects are standardized. Standard errors are in parentheses. Oral reading fluency was the only timed test at 60 seconds.
*p < .10. **p < .05. ***p < .01.

6.2. Analysis of Monitoring Data

As described in **Section 5.2**, AIR researchers also analyzed reports of teacher commentary that ARED collected as part of program monitoring to supplement the quantitative findings. These are self-reported anecdotal information reported by teachers and were only mentioned by less than 1-7% in each instance. Common issues that emerged from these data were the following: perceived large number of students making it difficult to manage the class, high rates of student absences, missing or absent necessary materials for instruction, inaccuracies in materials, and difficult to use numeracy materials. Given the small number of instances these were raised in, we mention them here briefly to complement the impact results and to take them into consideration for scaling *Ndaw Wune* programming.

6.3. Cost Analysis

Based on the cost analysis methodology described in Section 4.3, we estimated the costs of implementing *Ndaw Wune Année Scolaire*, which we present as follows, by evaluation question. We present all costs in both West African francs (FCFA) and United States dollars (USD), using a conversion rate of 580 FCFA per 1 USD.

Ndaw Wune Année Scolaire

We estimated the overall costs of the 2024–2025 *Année Scolaire* program to be 251,141,271 FCFA (USD 433,002) and the direct cost of implementation to be 226,027,144 FCFA (75,342 USD).[36] These costs cover 65 schools across six IEFs, including 100 classes and 3,000 total pupils. The overall cost per pupil is 83,714 FCFA (USD 144), and the direct cost per pupil is 389,702 FCFA (130 USD).

Exhibit 2021. Total Costs of Année Scolaire

Overall cost	Cost (FCFA)	Cost (USD)
Total cost	251,141,271	433,002
Cost per student	83,714	144

Our cost estimate is substantially lower than the ARED annual budget for *Année Scolaire* (544,843 USD), a common finding when comparing actual spending against budgeted spending. The difference may be attributable to small adjustments in how the program is actually implemented, as well as assumptions made in the cost analysis. For example, the cost analysis uses the current G50 protocol to price travel expenses, which has lower reimbursement rates than the G50 protocol that was in place for the majority of the 2024-2025 school year.

Next, we investigate costs by program component. To do so, we divided the 2024–2025 *Année Scolaire* program into the following four components:

1. Initial teacher training
2. Direct instruction
3. Pedagogical days
4. Management, planning, and evaluation

Initial teacher training. Before the start of the program each year, ARED conducts a training for all the teachers who will be teaching the *Ndaw Wune* students in the classroom. This

[36] Direct costs of implementation are overall costs minus the 10% overhead cost.

training consists of a “harmonization day” for all the teacher trainers in a central location (Diourbel for the 2024–2025 school year), followed by a travel day and then four days of training in each IEF. Costs include personnel time for ARED staff, consultants, and inspectors, transportation costs and per diem rates for meals and lodging, the annualized cost of projectors and laptops, and overhead costs estimated at 10% by ARED.

Direct instruction. In the “direct instruction” category, we include the costs for implementing the *Ndaw Wune* program in the classroom and ongoing program monitoring. ARED’s monitoring practices during SY2024-25 included regular reporting from supervisors and subsequent reports by ARED staff. Costs include personnel costs for ARED staff; stipends for teachers, supervisors, and inspectors; and printing and delivery costs for all educational materials. Facility costs are not included, as they are not a cost to the program, and overhead costs are estimated at 10%.

Pedagogical days. Twice each school year, ARED hosts “pedagogical days,” or refresher trainings, for teachers participating in the program. As part of the pedagogical days, ARED staff travel to each of the IEFs to observe classrooms and meet with school directors. Then, they host refresher trainings with participation from teachers, supervisors, focal points, and school directors. In the 2024–2025 school year, ARED organized the pedagogical days into three groups (one including the Kaolack and Bambey IEFs, one including the Saint-Louis and Podor IEFs, and one including the Kanel and Matam IEFs). Our cost estimates for the pedagogical days include personnel time for ARED staff, facilitators, and inspectors, transportation costs and per diem rates for meals and lodging, and overhead costs estimated at 10%.

Management, planning, and evaluation. In this category, we account for all remaining program activities that take place in the 3 months of the school year when *Année Scolaire* is not in session, including administrative work to manage the program and plan for program activities, as well as the annual evaluation costs. Evaluation costs include one day of enumerator training in a central location per year and two days of data collection in each IEF per round. Within this category, we include personnel costs for ARED staff and a stipend of 30,000 FCFA per enumerator per day of data collection, transportation costs and per diem rates for meals and lodging, annualized costs of tablets and a subscription to Survey CTO software, facility costs for renting the room for enumerator training, and overhead costs estimated at 10%.

Exhibit 21 reports the costs of each program component and the percentage of total program costs allocated to each component. Direct instruction accounts for the largest portion of total program costs (~62%), followed by management, planning, and evaluation (~23%). The two rounds of pedagogical days account for about 7%, and the week of teacher training is about 8% of the total costs.

Exhibit 2122. Costs of Année Scolaire by Program Component

Component	Cost (FCFA)	Cost (USD)	% of total
Initial teacher training	19,867,653	34,255	7.91%
Direct instruction	155,449,958	268,017	61.90%
Pedagogical days	18,713,517	32,265	7.45%

Management, planning, and evaluation	57,110,144	98,466	22.74%

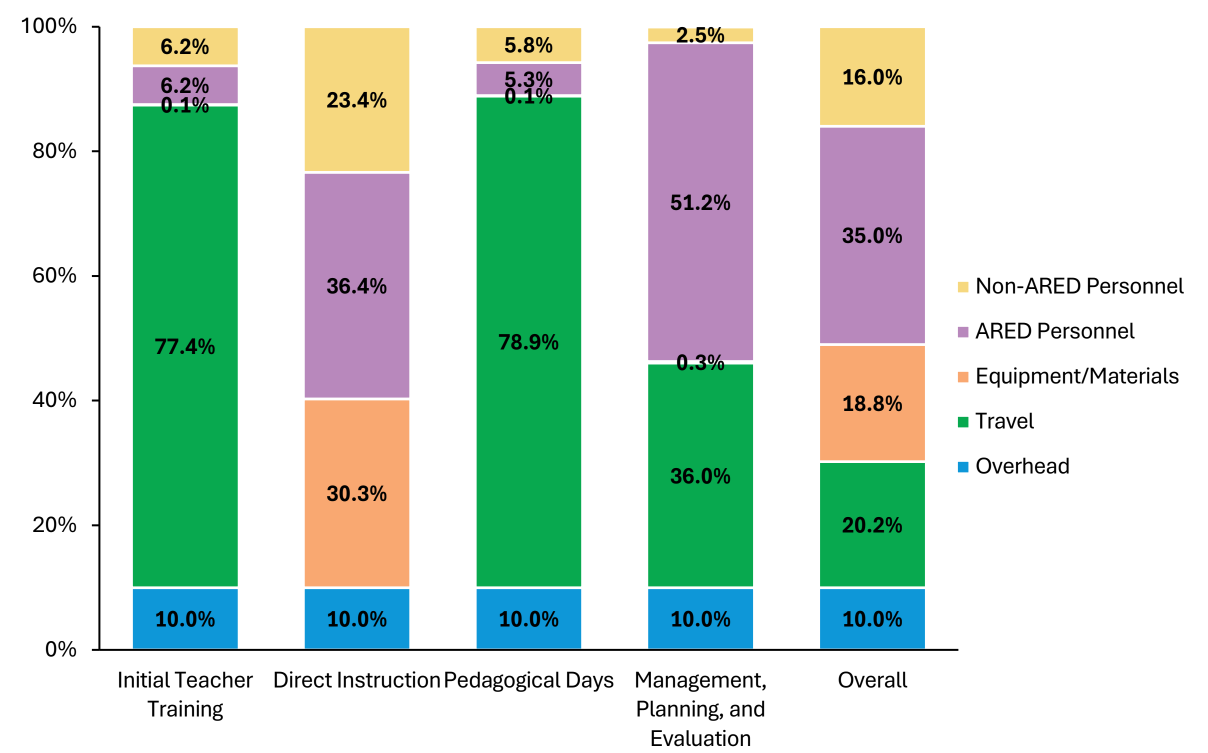
We analyze the total costs by resource category in **Exhibit 22**. The largest allocation of resources goes toward personnel costs, with ARED salaries accounting for the majority of the personnel category. Travel costs make up the majority of non-personnel costs. Overhead costs are assumed to be 10%, based on conversations with ARED staff.

Exhibit 232. Costs of *Année Scolaire* by Resource Category

Resource category	Cost (FCFA)	Cost (USD)	% of total
Personnel	128,065,591	220,803	50.99%
ARED personnel	88,005,591	151,734	35.04%
Consultants/Facilitators	1,800,000	3,103	0.72%
Teachers	30,000,000	51,724	11.95%
Supervisors	6,000,000	10,345	2.39%
Inspectors	820,000	1,414	0.33%
Evaluators	1,440,000	2,483	0.33%
Non-personnel	97,961,553	168,899	39.01%
Travel	50,705,099	87,423	20.19%
Equipment/Materials	47,231,454	81,434	18.81%
Facility	25,000	43	0.01%
Overhead	25,114,127	43,300	10.00%
Total	**251,141,271**	**433,002**	**100.00%**

In **Exhibit 23**, we analyze the resource allocation within each program component, compared to the overall allocation of costs for the program. Travel costs make up the majority of costs for the initial teacher training and the pedagogical days, while personnel costs account for most expenditures in the direct instruction and management, planning, and evaluation components. A full table of costs by program component and resource category can be found in **Appendix C**.

Exhibit 2324. Allocation of Costs for *Année Scolaire* by Program Component



Because ARED pays teachers, supervisors, and inspectors the same stipend rates across all regions and the G50 protocol stipulates rates for travel reimbursement and stipends for government personnel that are similar nationwide, we do not find substantive variations by program location. The primary differences include transportation costs, which increase for ARED staff based on the distance of implementation sites from Dakar, and, in some cases, printing costs. ARED reported that some materials (e.g., teacher guides and student math workbooks) have the same printing costs across languages, while others (e.g., student reading workbooks) vary by language, due to differences in the number of pages by language.

Scaling Scenarios

Lastly, we present cost projections for four scaling scenarios for *Ndaw Wune Année Scolaire.* For each scenario, the program is scaled to 50,000 students in 1,670 classes across the same six IEFs as the base scenario (Podor, Matam, Saint-Louis, Kaolack, Bambey, and Kanel). Under the scaling scenarios, ARED adds two salaried personnel, per-item materials costs are reduced according to World Bank databases for large-scale printing, all evaluation costs are removed, and all personnel costs associated with the development of educational materials are also removed.

- Scenario 1: Scale to 50,000 students in the six IEFs, with no change to the number of hours of instruction, initial teacher training, or pedagogical days. School directors assume the role of supervisors.
- Scenario 2: Scale to 50,000 students in the six IEFs, with no change to the number of hours of instruction. The initial teacher training is reduced to four half-days. Stipends are removed for supervisors and inspectors.

- Scenario 3: Scale to 50,000 students in the six IEFs, with all instruction occurring during official school hours, for a maximum of one hour for math and one hour for literacy per week. The initial teacher training is reduced to four half-days. Stipends are removed for supervisors, inspectors, and teachers.
- Scenario 4: All the assumptions of Scenario 3, and students use existing materials such that all costs for the printing and delivery of instructional materials are removed.

Scenario 1. In the first scenario, teacher stipends are reduced to 25,000 FCFA per month. The supervisor role is assumed by school directors, at a rate of 65 supervisors per every 100 schools, for a stipend of 30,000 FCFA per trimester. A total of 112 facilitators for teacher trainings and pedagogical days are needed to train 1,670 teachers in 56 districts, and they are compensated 10,000 FCFA per day. Because trainings occur in each district, non-resident travel costs are replaced with resident travel costs. It is assumed that one laptop and one projector are required for each training (i.e., one per district). As shown in **Exhibit 24**, the estimated cost for Scenario 1 is 767,650,752 FCFA (1,323,536 USD). The per-student total cost is 15,353 FCFA (26 USD), and per-student direct costs are 13,818 FCFA (24 USD).

Exhibit 254. Estimated Costs for Scenario 1

Resource category	Cost (FCFA)	Cost (USD)
Personnel	377,768,617	651,325
ARED personnel	87,148,617	150,256
Facilitators	6,720,000	11,586
Teachers	250,500,000	431,897
Supervisors	32,580,000	56,172
Inspectors	820,000	1,414
Non-personnel	313,117,060	539,857
Travel	225,686,026	389,114
Equipment/Materials	87,431,035	150,743
Overhead	76,765,075	132,354
Total	**767,650,752**	**1,323,536**
Per-student	**15,353**	**26**

Scenario 2. In the second scenario, we keep many of the assumptions of Scenario 1 but reduce the initial teacher training to four half-days from four full days. We also remove the stipends for inspectors and supervisors; however, inspectors/focal points maintain a daily compensation of 10,000 FCFA to attend trainings. As shown in **Exhibit 25**, the estimated cost for Scenario 2 is 638,388,975 FCFA (1,100,671 USD). The per-student total cost is 12,768 FCFA (22 USD), and the per-student direct costs are 11,491 FCFA (20 USD).

Exhibit 2526. Estimated Costs for Scenario 2

Resource category	Cost (FCFA)	Cost (USD)
Personnel	344,888,617	594,636
ARED personnel	87,148,617	150,256
Facilitators	6,720,000	11,586
Teachers	250,500,000	431,897
Supervisors	0	0
Inspectors	520,000	897
Non-personnel	229,661,460	395,968
Travel	142,230,426	245,225
Equipment/Materials	87,431,035	150,743
Overhead	63,838,897	110,067
Total	**638,388,975**	**1,100,671**
Per-student	**12,768**	**22**

Scenario 3. In this scenario, we assume that all instruction occurs during regular school hours and we remove the teacher stipend and reduce the hours of instruction from six to a maximum of two. As shown in **Exhibit 26**, the estimated cost for Scenario 3 is 360,055,641 FCFA (620,786 USD). The per-student total cost is 7,201 FCFA (12 USD), and the per-student direct costs are 6,481 FCFA (11 USD).

Exhibit 276. Estimated Costs for Scenario 3

Resource category	Cost (FCFA)	Cost (USD)
Personnel	94,388,617	162,739
ARED personnel	87,148,617	150,256
Facilitators	6,720,000	11,586
Teachers	0	0
Supervisors	0	0
Inspectors	520,000	897
Non-personnel	229,661,460	395,968
Travel	142,230,426	245,225
Equipment/Materials	87,431,035	150,743
Overhead	36,005,564	62,079
Total	**360,055,641**	**620,786**
Per-student	**7,201**	**12**

Scenario 4. In this scenario, students use existing materials such that there are no cost for printing/delivering instructional materials. **Exhibit 27** reveals that removing costs for student materials further reduces the overall costs to 263,343,808 FCFA (454,041 USD) and per-student costs to 5,267 FCFA (9 USD). Costs of direct implementation are reduced to 237,009,427 FCFA (4008,637 USD) and 4,740 FCFA (8 USD) per student.

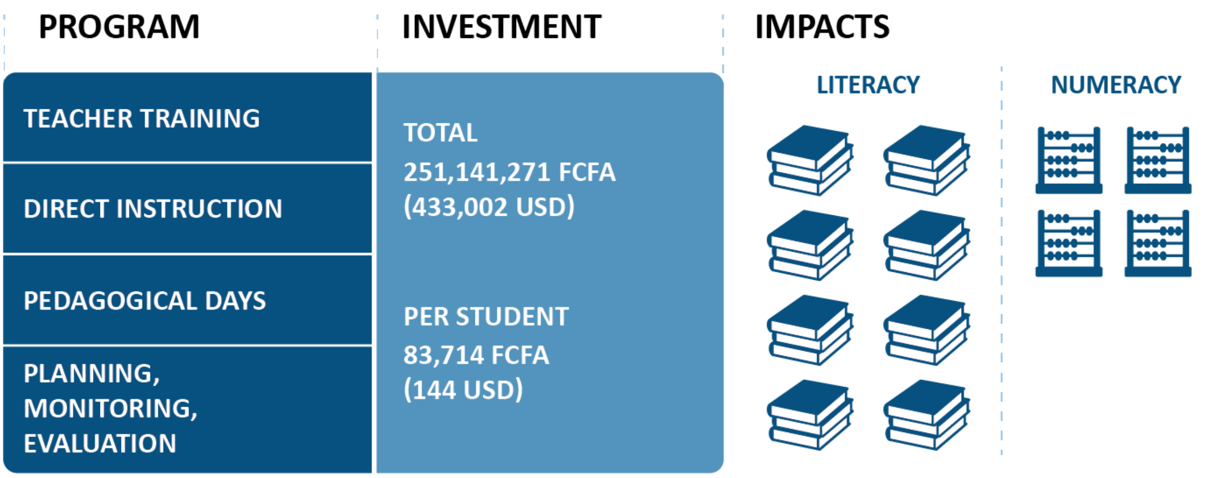
Exhibit 287. Estimated Costs for Scenario 4

Resource category	Cost (FCFA)	Cost (USD)
Personnel	94,388,617	162,739
ARED personnel	87,148,617	150,256
Facilitators	6,720,000	11,586
Teachers	0	0
Supervisors	0	0
Inspectors	520,000	897
Non-personnel	142,620,810	245,898
Travel	142,230,426	245,225
Equipment/Materials	390,385	673
Overhead	26,334,381	45,404
Total	**263,343,808**	**454,041**
Per-student	**5,267**	**9**

6.4. Cost-Effectiveness Analysis

Pairing the findings from the cost analysis of *Ndaw Wune Année Scolaire* with findings from the impact evaluation, we provide estimates of cost-effectiveness, using an impact inventory approach. Overall, we find that for a cost of 83,714 FCFA (or 144 USD) per student, *Ndaw Wune Année Scolaire* produced positive impacts on multiple foundational literacy and numeracy outcomes, as illustrated in **Exhibit 28**.

Exhibit 298. Illustrative Impact Inventory



In **Exhibit 29**, we linearly scale the effect sizes for each outcome to reflect the impact per $100 USD/student investment in *Ndaw Wune Année Scolaire*. We report the impact per $100 to facilitate comparisons of cost-effectiveness across multiple remedial education programs. As is common with these analyses, it is important to note that the ratio does not suggest that impacts scale linearly with investment. The investment of 58,000 FCFA (100 USD) per student produces moderate effect size increases in letter sound knowledge, word decoding, and oral reading fluency, as well as small effect size increases in reading comprehension and numeracy outcomes.

Exhibit 309. Impacts Achieved Per $100 Student Investment

Competency	Outcome (*SD* and 90% CI)
Letter knowledge and word decoding	0.29 SD (0.20–0.37)
Oral reading fluency	0.22 SD (0.09–0.35)
Reading comprehension	0.10 SD (0.01–0.19)
Number sense	0.13 SD (0.03–0.23)
Subtraction	0.14 SD (0.03–0.24)

7. Conclusions

Based on our evaluation, we find that *Ndaw Wune Année Scolaire* is an overall effective program for improving student foundational literacy scores in national languages for struggling students in Grades 2-3. Specifically, *Ndaw Wune* had significant positive impacts on word recognition sub-skills and reading comprehension skills in national languages. No significant impacts were found on phonological awareness and oral language competencies. *Ndaw Wune* had a smaller, but still positive, impact on students' foundational numeracy outcomes. We will discuss each of these in turn below within the theoretical framework of learning to read in multilingual Senegal.

To reiterate, CFRA theory (presented in Section 2) posits that foundational literacy is dependent on *both* word recognition sub-skills and language-comprehension sub-skills – working together with automaticity leading to reading comprehension skills. Treatment students in our study had significant positive results in several word recognition sub-skills, including letter sound knowledge (grapho-phonological processing) (0.45 *SD*), decoding skills (0.42 *SD*), oral reading fluency (0.32 *SD*), and reading comprehension (0.14 *SD*). This indicates that the *Ndaw Wune* program is very effective in building the components for the word recognition competencies needed for reading comprehension. Phonological awareness increased in magnitude, but this was not statistically significant; and there were no significant impacts on listening comprehension. Considering the context, where less than a third of students have access to learning and reading in a language they understand and speak well – mostly French (Nakamura & Holla, 2024), This is a powerful indicator that, when systematic instruction is provided on the decoding and word recognition sub-skills, reading comprehension can be improved when children are taught in languages they use and understand. Critically, *Ndaw Wune'*s success in fostering reading comprehension is likely due to the combination of effective structured pedagogy for word recognition (i.e., explicit phonics instruction, clear scope and sequencing of materials, assessment informed instructions, and efficient availability and linkages between student textbooks and teacher guides) *and* teaching children in languages they use and understand. These results highlight that the *Ndaw Wune* program **is successfully fostering foundational literacy skills in Senegalese Grade 2-3 children.**

Two of the six numeracy subtests showed positive and significant impacts, including number sense (0.19 *SD*) and subtraction (0.20 *SD*). All other subtests had no significant finding. The significant gains in number sense and subtraction are particularly meaningful because these skills form the foundation of mathematical understanding in early grades. Improved number sense indicates that students are better able to recognize patterns, understand numerical relationships, and solve problems involving missing numbers, skills that are essential for all future math learning. The positive impact on subtraction demonstrates progress in students' ability to perform basic operations, which is a critical step toward mastering more complex arithmetic. However, four numeracy subtests, counting, digit discrimination, addition, and number identification, did not show statistically significant gains. This suggests that the intervention was **effective in strengthening foundational mathematical reasoning and basic operations but may require further refinement to address the full spectrum of early grade math skills.** The mixed results underscore the importance of tailoring instructional strategies and materials to ensure comprehensive development across all key numeracy domains.

Subgroup impacts. As noted previously, the subgroup analysis was not powered to detect significant effects. We present suggestive conclusions here based on the data.

1. We find that both boys' and girls' outcomes were positively impacted by *Ndaw Wune*. Boys saw positive and significant impacts in three literacy subtests compared to two for girls, with the difference being reading comprehension. Boys' results were also consistently higher in magnitude. The results are more mixed for numeracy, but boys had positive treatment effects on the addition and subtraction sections, while girls had a positive treatment effect on number sense and a negative one on the counting subtest. Related studies have found mixed gendered effects depending on the context. For instance, in a study of a remedial literacy program in South Africa, the authors noted that girls outperformed boys at baseline and endline and that the girls' advantage held even in their multivariate regression models (Fleisch et al., 2017). The evaluation of a remedial program in Niger did not find any gendered differences in treatment effects (Maruyama & Kurosaki, 2021).
2. We did not have access to information about the language of instruction in schools before comparison school sampling was done. In the end, we had data for 12 Serer treatment schools and only one Serer comparison school. This was not a large enough sample to make reliable comparisons between the treatment and comparison groups. The sample included 12 Wolof treatment and 23 Wolof comparison schools and 11 Pulaar treatment and comparison schools. Students who

speak Wolof saw the largest gains. The data indicate that Wolof speakers generally had stronger treatment effects on the literacy subtests than Pulaar speakers. Although Wolof-speaking treatment students showed a negative treatment effect on listening comprehension, they demonstrated stronger gains in decoding skills (0.49 *SD* vs. 0.33 *SD* for Pulaar speakers) and achieved a 0.53 *SD* treatment effect on oral reading fluency, while the treatment effect for Pulaar speakers in oral reading fluency was not statistically significant.

3. Regionally, the strongest results came from Kaolack. In Kaolack, three of the five literacy subtests were significantly positive, compared to one for Saint-Louis and none for Matam. None of the individual regions had a positive and significant treatment effect in the numeracy section.

Our findings fit into the larger context of education evaluations from LMICs. A meta-analysis from 32 countries (67 studies) in LMICs showed that structured pedagogy programs had an overall effect of 0.30 *SD* on foundational literacy outcomes, with stronger effects on lower order word recognition sub-skills, and with impact differing with degree of teacher support and on the whether the program was in an first language (L1) or second language (L2) (Kim, Lee, & Zuilkowski, 2020). This skew towards word recognition sub-skills benefiting from early reading programs in LMICs is common and mirrored in this study's findings as well. In another example, an impact evaluation looking at a mix of interventions in Zambia found positive and significant impacts between 0.18 *SD* and 0.32 *SD* on EGRA subtests for decoding, oral reading fluency, and reading comprehension, along with impacts between 0.13 *SD* and 0.14 *SD* on EGMA subtests for number sense, addition, and subtraction (De Hoop et al., 2025). An evaluation of an early grade reading intervention in Liberia found effect sizes of ranging from 0.52 to 1.23 *SD* on letter sound knowledge, 0.55 on phonemic awareness, 0.78 on familiar word decoding, 1.23 *SD* on invented word decoding, 0.80 *SD* on oral reading fluency, and 0.82 *SD* on reading comprehension. The same study also found effect sizes between 0.14 *SD* and 0.34 *SD* for EGMA subtests despite the intervention not having a numeracy component (RTI, 2011). Finally, an evaluation of a remedial education program in Colombia which used an EGRA to measure student literacy found a 0.36 *SD* impact on letter sound knowledge and a 0.19 SD impact on oral reading fluency. It found no significant impact on invented word decoding or reading comprehension (Álvarez Marinelli, 2024).

Finally, we estimate the total cost of *Ndaw Wune Année Scolaire* to be **251,141,271 FCFA (433,002 USD)** or **83,714 FCFA (144 USD) per student**. We find that most of the cost is consumed by direct instruction activities (62%), followed by management, planning, and evaluation (23%). About half of program costs (51%) go to direct personnel costs, while nearly a fifth (20%) of costs are used to fund travel for trainings, pedagogical days, and evaluations. The **cost-effectiveness analysis** finds that a $100 investment per student produces moderate effect size increases in letter sound knowledge, word decoding, and oral reading fluency, as well as small effect size increases in reading comprehension and numeracy outcomes.[37]

[37] There are few cost analyses of education programming in sub-Saharan Africa, and those that exist tend to be for programs embedded within the existing education delivery structure (i.e., personnel costs are nonexistent or quite low). Thus, caution should be applied when comparing the cost of delivering the standalone *Ndaw Wune Année Scolaire* program to the costs of other education programs.

8. Recommendations for Further Increasing Impact

Finally, we present several recommendations for ARED, the Gates Foundation, and MEN to help improve programming in the future:

For Teacher Professional Development and Classroom Instruction:

1. **Expand explicit phonics instruction in regular classrooms**. The *Ndaw Wune* program demonstrated that supporting grapho-phonological mapping skills through explicit phonics instruction and providing structured opportunities for improving automaticity (speed and accuracy of making sound-symbol connections at the sub-lexical and lexical levels) are proving highly effective for improving decoding sub-skills and reading comprehension. Given the strong positive results for these sub-skills from the *Ndaw Wune* evaluation, we also recommend that MOHEBS draw from these findings and consider integrating additional explicit phonics instruction in Senegalese languages in classrooms across the country, which has not been happening in most classrooms (over 73%, Nakamura et al., 2024) in Senegal.
2. **Ensure reinforced implementation of the structured pedagogical elements for teaching foundational skills** in classrooms across Senegal. The combination of scope and sequence-based planning, clear and scripted (more so for language arts than for math) teacher guides, and student and teacher learning materials, especially workbooks for each child, have significant impact, as per the results of this evaluation. These aspects of instruction should be the core of pre-service teacher preparation programs and in-service training and tracked through implementation fidelity monitoring in foundational learning programs across the country. Given large class sizes, the lack of teaching and learning materials other than textbooks and teacher guides in MOHEBS classrooms, minimal exposure to print outside school, and chalkboards that are in poor condition in many classrooms, the importance of a leveled workbook for each child is a critical component of the *Ndaw Wune* approach.
3. **Consider further oral language support in national languages.** The results of the evaluation found no significant improvement in oral language skills – one of the necessary core components of foundation literacy development. Further, this result was not due to a ceiling effect in oral language skills at baseline. Therefore, *Ndaw Wune* and other programs nationally must bolster listening comprehension and oral vocabulary skills – even if already teaching in languages children know and understand. In other words, even when children are taught in their own language, there is a need to spend considerable teaching time on bolstering vocabulary and linguistic knowledge across the curriculum and not focus solely on phonics and decoding instruction.
4. **Invest in improvements for numeracy components of the program.** The impact evaluation revealed challenges with numeracy instruction. Therefore, we provide some suggested approaches for improving foundational numeracy instruction. First, instruction should follow clear progression within each operation, moving from viewing multi-digit numbers as strings of individual digits to understanding the full place-value structure (e.g., 10 ones = one 10) (Rogers, 2023). Second, teachers could explicitly teach efficient calculation strategies beyond tallying, such as bridging

through ten, jumping, and rounding and adjusting methods shown to significantly improve mental math fluency when modeled and practiced systematically (Kumm & Graven, 2024). Third, structured number talks could be incorporated to encourage students to share and compare multiple solution strategies, fostering number sense, flexibility, and metacognitive growth through collaborative discourse. Finally, extensive use of grouping materials, such as bundles of sticks or locally made base-ten blocks can make abstract place-value concepts tangible (e.g., reinforcing the idea that 10 ones form a ten) and support regrouping in operations (Resnik, 1987).

5. **Need for supporting transfer and French oral language skills given Senegal's bilingual education goals.** The goal of *Ndaw Wune* is improvement of literacy skills in national languages to give students the foundation they need to learn to read in a second language. The goal of foundational learning in Senegal is literacy ability in national languages and French (Nakamura et al., 2026). The science of bilingual reading makes it clear that learning to read in a new language is critically dependent on how well students can read in a first language, and how much oral (or signed) language proficiency they have in the new language they need to learn. More emphasis on the cognitive underpinnings of biliteracy acquisition would better support students' transfer of their L1 skills to an L2.

For Further Research:

6. **Understand the intersections between language and numeracy development.** Future research and program design in Senegal should systematically investigate how language skills and the choice of language of instruction shape students' numeracy development. Research should explore how bilingual or multilingual teaching, where students learn math in both their home language and French, affects mathematical reasoning, transfer of skills, and long-term achievement. For example, does early mastery of numeracy in Wolof or Pulaar facilitate later learning of math in French, or does it create challenges in transitioning mathematical vocabulary and concepts? Future research should also study how language mismatches affect numeracy outcomes over time.
7. **Ensure data of matching teacher/student language is collected.** With the advent of large-scale initiatives in national languages, the MEN plans to collect teacher language data more systematically in the future within the MIRADOR (information system for teacher management) to support the implementation of bilingual education. We recommend supplementing this with student language data to understand the extent of student-teacher language match/mismatch.

8. **Carry out a process evaluation to delve into the elements of the program that most likely lead to improved student achievement and the reasons for the regional variations.** As ARED seeks to refine and scale-up *Ndaw Wune,* a corresponding process evaluation would be tremendously useful in identifying key elements of the program that are helpful for students and teachers and noting how the program can replicate the successes and address any issues in the program to improve it for scale-up. Such an evaluation could examine the fidelity of implementation by comparing intended activities such as structured lesson delivery and use of instructional materials with what actually occurred across schools. It could also assess the quality

of delivery through classroom observations that capture how well teachers apply the pedagogical methods promoted by the program, the consistency of coaching support, and the nature of any adaptations made in practice. Gathering teacher perspectives through interviews, surveys, and focus groups could help illuminate how teachers perceive the usefulness of the training and materials, the challenges they face, and the conditions that most enable or hinder effective implementation. The evaluation could also look closely at student engagement and classroom dynamics to identify the instructional practices that most consistently lead to active participation and comprehension. In addition, tracking program dosage (i.e., how much exposure students actually have to *Ndaw Wune* lessons) could help explain variation in effects across schools and regions.

9. **Systematically track student attendance data.** Since the attendance data we received was inconsistent and had several gaps, future work should explore instituting a systematic and cost-effective data tracker which can track individual student attendance. This would help in identifying reasons for absences and uncover chronic absenteeism. This can include a simple SMS-based tool that can feed into a data-driven decision-making tool and would not require internet or smartphones. Such an approach has been successfully piloted in several countries, including those in West Africa (Mugo & Piper, 2025). For instance, a pilot in Kenya found that the cost of maintaining an SMS-based attendance system for 10,000 learners at a central level, supplemented by weekly personalized SMS to teachers on struggling learners would cost less than 0.08 USD per child per year (Mugo & Piper, 2025). Improved data systems will also help track attendance, attrition, and heterogeneous impacts, enabling more responsive program management.

10. **Utilize artificial intelligence (AI) and education technologies (where appropriate) to assist with provision of additional teaching and learning materials (TLMs) and trainings.** AI-generated short videos could be used for frequent light-touch training sessions and supportive coaching. It could also be a supplementary input on tough topics or practice sessions on numeracy-related topics. It may also be useful to utilize AI to produce additional "back-up" materials in line with all the science of bilingual reading principles discussed above, with the caveat that AI translation and content generation are not yet available for Senegalese languages. Some examples include:

    a. Assurance that every student has access to textbooks/workbooks, every teacher has a guide and they are aligned. In rare cases, where students need to share books, teachers should be guided as to how to guide students through sharing activities (part of the teacher training).

    b. Utilizing AI to create pictures that can be used as story cards, letter cards, and other TLMs. These can be created much more cost-efficiently and are not always reliant on current linguistic constraints of AI.

    c. Ensuring materials within the student textbook address all literacy and language sub-skills in an appropriate learning trajectory, especially for biliteracy acquisition and later French transition, for there to be fewer pieces to print and align.

For Policy:

11. **Importance of initial reading instruction in Senegalese languages**. The evaluation of the *Ndaw Wune* program makes it clear that when teaching children in languages they use and understand, with a focus on explicit teaching of letter sounds and phonics rules, significant improvements occur in students decoding, and reading comprehension, even for struggling, high risk students. *Ndaw Wune* is an exceptional example of how, when students are taught in languages they understand, resources can be focused on the word recognition aspect of foundational literacy. This approach, as opposed to teaching children in a foreign language, makes literacy acquisition much more efficient as precious classroom time is not wasted on teaching introductory oral language skills, and focuses on letter sounds, blending, fluency, and higher-level oral language skills.
    In addition to the use of a language students know, it is also vital for teachers to know the language of instruction. The data from this study did not find differences between student language knowledge and the language of instruction, but we did not collect data on teacher language knowledge. There are still considerable mismatches across the Senegal between students, teachers and the language of instruction in particular schools, especially for smaller languages.[38,39] These need to be addressed for effective foundational learning and numeracy bilingual instruction.

[38] https://www.air.org/sites/default/files/2025-08/BUILD-Introduction-to-the-Framework-August-2025.pdf
[39] https://scienceofteaching.site/wp-content/uploads/2025/08/English-Implications-of-and-Strategies-for-Dealing-with-Language-of-Instruction-Mismatch-in-Senegal_Final.pdf

References

Ahouakan Ehouman, W. V., & Diene, M. (2017). Does school quality matter? Primary schools' characteristics and child labor intensity in Senegal. *Journal of African Development, 19*(1), 113–131. https://doi.org/10.5325/jafrideve.19.1.0113

All children in Senegal are born to learn. (n.d.). https://www.unesco.org/gem-report/sites/default/files/medias/fichiers/2023/02/Spotlight%20Advocacy_Senegal_EN_v1.pdf

Álvarez Marinelli, H. A., Berlinski, S., & Busso, M. (2024). Remedial education: Evidence from a sequence of experiments in Colombia. *Journal of Human Resources, 59*(1), 141–174. https://doi.org/10.3368/jhr.0320-10801R2

Angrist, N., Meager, R., Angrist, & London School of Economics. (2023). Implementation matters: Generalizing treatment effects in education. In *EdWorkingPaper* (No. 23–802). https://edworkingpapers.com/sites/default/files/ai23-802.pdf

Bagby, E., Swift-Morgan, J., Niang, A., Upadhyay, A., & Chemonics International Inc. (2022). Achieving Cost-Effective Instructional Coaching at Scale: Evidence from Senegal. In *Learning From All Children Reading*. https://chemonics.com/wp-content/uploads/2022/04/ACR_Tech_Brief_LPT_Coaching_CIES.pdf

Banerjee, A. V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2016, November 20). *Mainstreaming an* Effective *effective intervention: Evidence from* Randomized *randomized evaluations of "Teaching at the Right Level" in India*. (MIT Department of Economics Working Paper No. 16-08). SSRN. https://doi.org/10.2139/ssrn.2846971

Brander, R. L., Puett, C., Becquey, E., Leroy, J. L., Ruel, M. T., Sessou, F. E., & Huybregts, L. (2024). The cost and cost-effectiveness of an integrated wasting prevention and screening intervention package in Burkina Faso and Mali. *The Journal of Nutrition, 154*(8), 2551–2565. https://doi.org/10.1016/j.tjnut.2024.04.010

Bronfenbrenner, U. (2000). Ecological systems theory. In A. E. Kazdin (Ed.), *Encyclopedia of Psychology* (Vol. 3, pp. 129–133). Oxford University Press. https://doi.org/10.1037/10518-046

Brown, L. E., Kim, H. Y., Dolan, C. T., Brown, A., Sklar, J., & Aber, J. L. (2022). Remedial Programming and Skill-Targeted SEL in Low-Income and Crisis-Affected contexts: Experimental evidence from Niger. *Journal of Research on Educational Effectiveness*, *16*(4), 583–614. https://doi.org/10.1080/19345747.2022.2139785

Cost-effective approaches cost-effective approaches to improve global to improve global learning. (2023). *What does recent evidence tell us are "Smart Buys" for improving learning in low-and middle-income countries? Recommendations of the Global Education Evidence Advisory Panel (GEEAP)*. https://thedocs.worldbank.org/en/doc/231d98251cf326922518be0cbe306fdc-0200022023/related/GEEAP-Report-Smart-Buys-2023-final.pdf

Danks, A., & Kendziora, K. (2021, April). *Standards for the economic evaluation of educational and social programs*. American Institutes for Research. https://www.air.org/sites/default/files/Standards-for-the-Economic-Evaluation-of-Educational-and-Social-Programs-CASP-May-2021.pdf

De Hoop, T., Ring, H., Siwach, G., Dias, P., Tembo, G., Taube, A., & Behmer, T. (2025, January). *Mixed-methods cluster-randomized controlled trial of Impact Network's eSchool 360 model in rural Zambia: Endline report*. American Institutes for Research. https://www.air.org/sites/default/files/2025-01/Endline-Report-Impact-Network-eSchool-RCT-Trial-360-Zambia-Jan-2025.pdf

Duflo, A., Kiessel, J., & Lucas, A. (2020). *Experimental evidence on alternative policies to increase learning at scale*. https://doi.org/10.3386/w27298

Duflo, Annie, et al. "Experimental Evidence on Four Policies to Increase Learning at Scale." *the Economic Journal/Economic Journal*, 2 Feb. 2024, https://doi.org/10.1093/ej/ueae003.

Encinar, Pierre-Emmanuel. (2022). PASEC 2019 study report in English

Evans, D., & Mendez Acosta, A. (2020). *Education in Africa: What Are We Learning?* https://www.cgdev.org/sites/default/files/education-africa-what-are-we-learning.pdf

Evans, D. K., & Mendez Acosta, A. (2021). Education in Africa: What are we learning? *Journal of African Economies, 30*(1), 13–54. https://doi.org/10.1093/jae/ejaa009

Fleisch, B., Taylor, S., Schöer, V., & Mabogoane, T. (2017). Failing to catch up in reading in the middle years: The findings of the impact evaluation of the Reading Catch-Up Programme in South Africa. *International Journal of Educational Development, 53*, 36–47. https://doi.org/10.1016/j.ijedudev.2016.11.008

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement, 33(3), 613–619.

Fuson, K. C., & Briars, D. J. (1990). Using a Base-Ten Blocks Learning/Teaching Approach for First- and Second-Grade Place-Value and Multidigit Addition and Subtraction. *Journal for Research in Mathematics Education*, *21*(3), 180. https://doi.org/10.2307/749373

Greenwood, C. R., Carta, J. J., Walker, D., Watson-Thompson, J., Gilkerson, J., Larson, A. L., & Schnitz, A. (2017). Conceptualizing a public health prevention intervention for bridging the 30 million word gap. *Clinical child and family psychology review*, *20*(1), 3-24.

Harden, K., Punjabi, M., & Fernandez, M. (2022). Influences on Teachers' Use of the Prescribed Language of Instruction: Evidence from Four Language Groups in the Philippines. *Education Quarterly Reviews*, *5*(1). https://doi.org/10.31014/aior.1993.05.01.460

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American educator*, *27*(1), 4-9.

Hoover, W. A., & Tunmer, W. E. (2020). *The cognitive foundations of reading and its acquisition: A framework with applications connecting teaching and learning*. Springer International Publishing. https://link.springer.com/book/10.1007/978-3-030-44195-1

Jacob, B. A., & Lefgren, L. (2002). *The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago* (NBER Working Paper No. 8916). National Bureau of Economic Research. https://www.nber.org/papers/w8916

Kébé, A. (n.d.). *Revue ASTR-GHANA Language and Multilingualism in Senegal: From Colonial Legacies to Contemporary Educational Issues*. Retrieved October 1, 2025, from https://edition-efua.acaref.net/wp-content/uploads/sites/6/2025/08/1-Abou-Bakry-KEBE.pdf

Koda, K. (2008). Impacts of prior literacy experience on learning to read in a second language. In K. Koda & A. M. Zehler (Eds.), *Learning to read across languages: Cross-linguistic relationships in first- and second-language literacy development* (pp. 68–96). Routledge.

Kumm, M., & Graven, M. H. (2024). Exploring pre-service teachers' knowledge of efficient calculation strategies. *South African Journal of Childhood Education*, *14*(1). https://doi.org/10.4102/sajce.v14i1.1494

LARTES-IFAN. (2017). Jàngandoo, Baromètre de la qualité des apprentissages au Sénégal - Principaux résultats.

Levin, H. M., McEwan, P. J., Belfield, C. R., Bowden, A. B., & Shand, R. D. (2018). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis* (3rd ed.). SAGE Publications. https://searchworks-lb.stanford.edu/view/13785876

Maruyama, T., & Kurosaki, T. (2021). Do remedial activities using math workbooks improve student learning? Empirical evidence from scaled-up interventions in Niger. *World Development, 148*, Article 105659. https://doi.org/10.1016/j.worlddev.2021.105659

Maruyama, Takao and Igei, Kengo. (2022). Community-Wide Support for Primary Students to Improve Basic Reading and Math Learning: Empirical Evidence from Madagascar https://ssrn.com/abstract=4076787 or http://dx.doi.org/10.2139/ssrn.4076787

Mohohlwanea, N., Taylor, S., Cilliers, J., & Fleisch, B. (2024). Reading skills transfer best from home language to a second language: Policy lessons from two field experiments in South Africa. *Journal of Research on Educational Effectiveness, 17*(4), 687–710. https://doi-org.air.idm.oclc.org/10.1080/19345747.2023.2279123

Monitoring Impacts on Learning Outcomes MAIN REPORT. (n.d.). https://milo.uis.unesco.org/wp-content/uploads/sites/17/2022/01/MILO-Summary-Full-Report.pdf

Moussa, S., Cissé, R., and Lô, C. (2021). How remediation helps to improve children's educational learnings in primary education in Senegal. In I. I. Munene (Ed.), *Ensuring all children learn: Lessons from the South on what works in equity and inclusion*. Lexington Books.

Mugo, P., & Piper, B. (2025, January 14). *Using SMS as a low-cost means to systematically track chronic absenteeism in schools: Reflections from South Sudan and Sierra Leone*. UKFIET – The Education and Development Forum. https://www.ukfiet.org/2025/using-sms-as-a-low-cost-means-to-systematically-track-chronic-absenteeism-in-schools-reflections-from-south-sudan-and-sierra-leone/

Muyabi, P. (2017). *Parental involvement in reducing learner absenteeism in selected primary schools of Kazungula district in the Southern Province of Zambia*. https://dspace.unza.zm/items/0eb43fa2-469b-4b03-beee-60fb3078e940

Nag, S., Vagh, S. B., Dulay, K. M., & Snowling, M. J. (2019). Home language, school language and children's literacy attainments: A systematic review of evidence from low- and middle-income countries. *Review of Education, 7*(1), 91–150. https://doi.org/10.1002/rev3.3130

Nakamura, P., Molotsky, A., Zarzur, R. C., Ranjit, V., Haddad, Y., & De Hoop, T. (2023). Language of instruction in schools in low-and middle-income countries: A systematic review. *Campbell Systematic Reviews, 19*(4), Article e1351. https://doi.org/10.1002/cl2.1351

Nakamura, P. R. & Holla, C. U. (2024). Semantic fluency: Measuring oral language proficiency to enable informed decision making on Language of Instruction in six countries. Comparative and International Education Society (CIES). Miami, Florida.

Nakamura, P., McBride, A., Holla, C., Yessifou, N., & Woodrick, K. (2026). *Blueprint for the use of evidence in language of instruction (LoI) decisions (BUILD): Senegal report* (Unpublished report). Bill & Melinda Gates Foundation.

Nakamura, P., Woodrick, K., McBride, A., Holla, C., & Zaas, D. (n.d.). *Blueprint for the Use of Evidence in Language of Instruction (LoI) Decisions (BUILD) Framework Introduction. Retrieved January 27, 2026, from https://www.air.org/sites/default/files/2025-08/BUILD-Introduction-to-the-Framework-August-2025.pdf*

Ndaw Wune: An Accelerated Learning Program to Support Education in Senegal - ARED. (2024, November 21). Ared-Edu.org. https://ared-edu.org/en/ndaw-wune-an-accelerated-learning-program-to-support-senegalese-entrepreneurship/

Ndour, C. T. (2024). *Understanding school dropout in Senegal: Exploring the impact of governance factors* (SSRN Scholarly Paper No. 4867074). Social Science Research Network. https://doi.org/10.2139/ssrn.4867074

Nishanthi, R. (2020). Understanding of the importance of mother tongue learning. *International Journal of Trend in Scientific Research and Development, 5*(1), 77–80. https://www.ijtsrd.com/papers/ijtsrd35846.pdf

Ouellette, G., Beers, A. A not-so-simple view of reading: how oral vocabulary and visual-word recognition complicate the story. *Read Writ 23*, 189–208 (2010). https://doi.org/10.1007/s11145-008-9159-1

Özek, U. (2021). *The effects of middle school remediation on postsecondary success: Regression discontinuity evidence from Florida* (Working Paper No. 258-0921). National Center for Analysis of Longitudinal Data in Education Research (CALDER). https://eric.ed.gov/?id=ED616057

Pillai, S. (2025, June 23). *Using SMS as a low cost means to systematically track chronic absenteeism in schools: Reflections from South Sudan and Sierra Leone*. The Education and Development Forum (UKFIET). https://www.ukfiet.org/2025/using-sms-as-a-low-cost-means-to-systematically-track-chronic-absenteeism-in-schools-reflections-from-south-sudan-and-sierra-leone/

Piper, B., Ralaingita W., Mejia J., Dubeck, M. M., DeStefano J., Stern J., Jordan R., & Sitabkhan Y. (2020). Structured Pedagogy How-to Guides and Literature Review https://scienceofteaching.site/wp-content/uploads/2022/10/SP_Introduction.pdf

Piper, B. & Dubeck, M. M. (2024). Responding to the learning crisis: Structured pedagogy in sub-Saharan Africa. *International Journal of Educational Development, 109*, Article 103095. https://doi.org/10.1016/j.ijedudev.2024.103095

Resnick, L. B. (1987). Learning in school and out. *Educational Researcher, 16*(9), 13–20. https://www.jstor.org/stable/749373

Rogers, A. (2023). *Rethinking the number magnitude-based progression: An analysis of place value development in Years 3–6* (MERGA45 Proceedings). Mathematics Education Research Group of Australasia. https://files.eric.ed.gov/fulltext/ED631611.pdf

RTI International. (2011). *Information for accountability: Impact evaluation of EGRA and teacher training (English)* (Working Paper No. 72457). World Bank. http://documents.worldbank.org/curated/en/772611468057237703

Singh, A., Mauricio Romero, Muralidharan, K., Alison Andrew, Noam Angrist, Jishnu Das, Jean Dreze, Peter Hull, Ajinkya Keskar, Ben Piper, Justin Sandefur, Ramamurthy Sripada, Dwithiya Raghavan, Akanksha Soni, & Indira Tirumala. (2024). COVID-19 Learning loss and recovery: Panel data evidence from India. In *INST TECH AUTONOMO DE MEXICO* [Journal-article]. https://doi.org/10.3368/jhr.0723-13025R2

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., Jobse, H., Geelen, M., Pastorello, M., and Eyers, J. (2015, December). *Interventions for improving learning outcomes and access to education in low- and middle- income countries: A systematic review* (3ie Systematic Review 24). International Initiative for Impact Evaluation. https://doi.org/10.23846/SRS007

Stern, J., Jukes, M., Piper, B., DeStefano, J., Mejia, J., Dubeck, P., Carrol, B., Jordan, R., Gatuyu, C., Nduku, T., Punjabi, M., Harris Van Keuren, C., & Tufail, F. (2021, October). *Learning at scale: Interim report*. RTI International. https://ierc-publicfiles.s3.amazonaws.com/public/resources/Learning%20at%20Scale%20Interim%20Report%20-%20Final%20Draft.pdf

Stern, J., Jukes, M. C. H., Cilliers, J., Fleisch, B., Taylor, S., & Mohohlwanea, N. (2024). Persistence and emergence of literacy skills: Long-term impacts of an effective early grade reading intervention in South Africa. *Journal of Research on Educational Effectiveness*. https://doi.org/10.1080/19345747.2024.2417288

Nakamura, P., Molotsky, A., Holla, C., McBride, A., & Gosselt, J. (2024). *Strengthening Bilingual and Multilingual Learning Systems in Francophone Africa Evidence from Senegal*. (n.d.). https://www.air.org/sites/default/files/2024-05/KIX-Strengthening-Bilingual-Education-Senegal-Brief-May-2024.pdf

Schwartz, A. (2012). *Remedial Education Programs to Accelerate Learning for All*. https://documents1.worldbank.org/curated/en/564671468151507990/pdf/797830WP0Anali0Box0379789B00PUBLIC0.pdf

Snilstveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., Jobse, H., Schmidt, T., & Jimenez, E. (2016). *Systematic Review Summary 7 The impact of education programmes on learning and school participation in low-and middle-income countries*. https://www.3ieimpact.org/sites/default/files/2019-05/srs7-education-report.pdf

Snilstveit, B., Gallagher, E., Phillips, D., Vojtkova, M., Eyers, J., Skaldiou, D., Stevenson, J., Bhavsar, A., & Davies, P. (2017). PROTOCOL: Interventions for improving learning outcomes and access to education in low- and middle-income countries: a systematic review. *Campbell Systematic Reviews*, *13*(1), 1–82. https://doi.org/10.1002/cl2.176

The Abdul Latif Jameel Poverty Action Lab (J-PAL). (n.d.). The Abdul Latif Jameel Poverty Action Lab (J-PAL). https://www.povertyactionlab.org/evaluation/impact-government-run-remedial-education-programs-student-learning-ghana

UNESCO Institute for Statistics. (2024). *Bulk Data Download Service*. https://uis.unesco.org/bdds

UNICEF. (2022). Teacher attendance and time on task in West and Central Africa. https://www.unicef.org/innocenti/media/4296/file/UNICEF-Time-to-Teach-Teacher-Attendance-WCA-2022.pdf

United States Agency for International Development (USAID). (2022). *Renforcement de la Lecture Initiale pour Tous (RELIT)*. https://www.rti.org/sites/default/files/relit_dec22_final_eng_approved.pdf

What Works Clearinghouse ™ Standards Handbook Version 4.0. (n.d.). https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

World Bank. (2024). *Senegal learning poverty brief - 2024 (English)*. World Bank Group. http://documents.worldbank.org/curated/en/099062624135013503

Appendix A. Sample Selection Test

Test Elève en Pulaar

Lecture

a	i	n	t	l	r	m	w	d	k
Demander à l'élève de lire les sons (pas les noms) des lettres. Merci de reporter la note de l'élève sur 10 à la feuille de notation									
na yo fof so mo to wi ko tan ne									
Demander à l'élève de lire les syllabes. Merci de reporter la note de l'élève sur 10 à la feuille de notation									
taro deftere wiyi nana wuro ladde gooto galle kala woni									
Demander à l'élève de lire les mots. Merci de reporter la note de l'élève sur 10 à la feuille de notation									

Mathématiques

11 10

19 32

7 5

16 13

18 29

Demander à l'élève de vous dire le chiffre le plus grand. Merci de reporter la note de l'élève sur 5

Appendix B. Program Locations

Année Scolaire (2024–2025)		*Vacances* (2024)	
IAs	**IEFs**	**IAs**	**IEFs**[40]
Diourbel	• Bambey	Diourbel	• Bambey • Diourbel • Mbacké
Kaolack	• Kaolack	Kaolack	• Nioro du Rip
Saint Louis	• Saint Louis • Podor	Matam	• Matam
Matam	• Matam • Kanel		

[40] These include six communes: Thilogne and Ourossogui in Matam, Ngogom in Bambey, Diourbel, Kael in Mbacké, and Paoskoto in Nioro.

Appendix C. Detailed Cost Methodology

In this appendix, we provide additional details on our costing methodology and highlight the assumptions we made regarding the cost analyses. On the following pages, we include tables that list the included costs in more detail.

ARED personnel costs. ARED provided salary data for 16 staff, along with the percentage of time devoted to *Ndaw Wune Année Scolaire* during the nine-month annual school year. We allocated ARED personnel costs across the different program components proportionately according to length of component and staff activity participation. For example, direct implementation of *Ndaw Wune Année Scolaire* consumes six of the nine months, while the management, planning, and evaluation component occupies the remaining three months; thus approximately 67% of ARED personnel costs are assigned to the direct implementation component and 33% are assigned to the management, planning, and evaluation component. For staff participating in the initial teacher training and/or pedagogical days, we allocate 2.56% of salary per week (assuming a nine-month salary spanning 39 weeks). For example, an ARED staff member who attended the initial teacher training and all pedagogical days would have their annual *Ndaw Wune Année Scolaire* salary allocated to each component as follows:

- 2.6% for initial teacher training
- 59% for direct instruction
- 5.1% for pedagogical days
- 33.3% for management, planning, and evaluation

Note that the allocations sum to 100% for all ARED personnel (thus full personnel costs are included in the overall program cost estimates), and parsing out personnel time by activity provides an approximate estimate to demonstrate the allocation of personnel costs for each activity.

Non-ARED personnel costs. ARED hired two consultants for the initial teacher training and one consultant for the pedagogical days; consultants were compensated 60,000 FCFA per day. Teachers, supervisors, and inspectors were provided a stipend for the six months of program implementation: Teachers were provided 50,000/month (for a total of 300,000), supervisors were provided 100,000/month (for a total of 600,000), and inspectors were provided 25,000/trimester (for a total of 50,000). Inspectors were paid an additional 10,000/day during the teacher training and pedagogical days. Evaluators were paid a stipend of 30,000 per person per day of data collection (for two two-day rounds total).

Materials and equipment. For the costs of materials, we used the printing and delivery costs that ARED incurred for *Ndaw Wune Année Scolaire.* For laptops and projectors, one of each was used in each IEF for training. We obtained average local prices and calculated the daily cost, for an estimated lifespan of six years for projectors and four years for laptops. The daily cost of the laptops and projectors was then apportioned to the teacher training and the pedagogical days, according to the number of days used.

Travel costs. We relied on conversations with ARED, budget documents, the TOR, and the G50 protocol to assign transportation, lodging, and meal costs.

For ARED staff transportation costs during the initial training, pedagogical days, and evaluation rounds, we followed the travel schedule in the TOR for the pedagogical days. The TOR divided the IEFs into three axes, with each axe containing two IEFs. For each axe, we assumed one driver and one vehicle, following the route described in the TOR (driving to the first IEF to drop off staff, then to the second, and then back to the first to pick up staff before returning to Dakar). We assumed a total daily rate of 95,000 FCFA for each driver/vehicle, which included driver pay, meals, and lodging costs. We assumed each vehicle achieved 8.5 km per liter of fuel and a fuel cost of 755 FCFA per liter.

For non-ARED staff transportation between regions, we calculated reimbursement costs as 75 FCFA per kilometer based on the distance between cities, as aligned with the G50 protocol and ARED's budget information. For transportation within regions, we calculated daily reimbursement rates for travelers who are residents of the village of the event as well as those traveling from outside the village, assuming that for events held in schools, one teacher and one director would be residents of the village and all other attendees would be residents from other areas in the region. Local reimbursement rates were 5,000 FCFA for residents and 10,000 for non-residents. For the pedagogical days, we assumed the 10,000 FCFA per diem for all attendees, per conversations with ARED.

For lodging, we followed the G50 protocol amount of 30,000 FCFA per night. We only calculated lodging costs for travelers from outside the region of the event. We assumed that ARED staff traveled the day before and the day after events when attending events in another region.

For meals, we calculated reimbursement rates for all travelers depending on whether they were residents of the region (in which case they received 6,000 FCFA reimbursement costs for two meals per day, according to the G50 protocol) or not (in which case they received 10,000 FCFA reimbursement costs for three meals per day).

Overhead. After consulting with ARED about overhead costs, we assigned overhead costs at a flat rate of 10% for each activity.

Initial teacher training. For the initial teacher training, we relied on the TOR for *Ndaw Wune Année Scolaire*, supplemented with information from interviews with ARED staff. The TOR and provided budget document accounted for five ARED staff, two paid consultants, four additional facilitators, six model teachers, 10 supervisors, and 10 focal points/inspectors to attend the harmonization day before the training began. For the four days of teacher training, we accounted for the aforementioned persons, plus 100 teachers and 66 directors (directors only attending the last day).

Pedagogical days. For the pedagogical days, we relied on the TOR for *Ndaw Wune Année Scolaire*. For local per diems, we assumed that one teacher, one director, and one supervisor resided in the village of the event for each IEF, with the rest residing outside the village. The TOR for *Ndaw Wune Année Scolaire* included three ARED staff, one paid consultant, three additional facilitators, 100 teachers, 10 supervisors, 12 IEF focal points, four IA focal points, and 56 directors.

Management, planning, and monitoring & evaluation. For this category, we relied on budget documents and communications with ARED. For the *Année Scolaire* program, we assumed two rounds of evaluation per year. Annual costs include one day of evaluator training (attended by four evaluators per IEF) and two days of data collection per round (conducted by two evaluators per IEF). We included the annualized costs of tablets to support data collection for the days of use, based on the assumption that each tablet would have four years of useful life. We also included annualized costs of SurveyCTO software, which ARED's monitoring & evaluation officer believed they would use in the future.

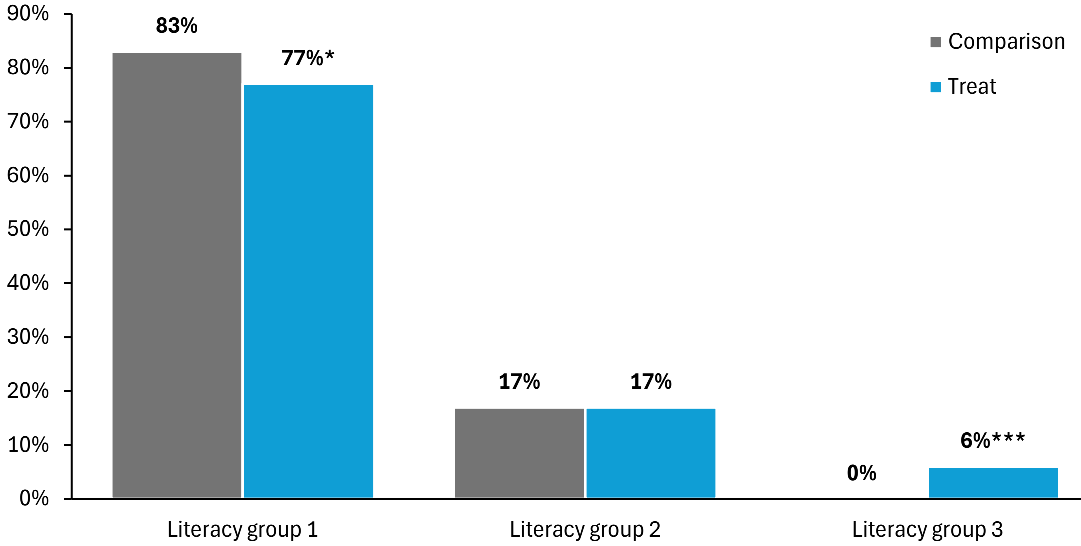
Exhibit C1. Summary of Annual Costs for *Ndaw Wune Année Scolaire*

Resource category	Cost (FCFA)	Cost (USD)
Initial teacher training		
Personnel	2,478,417	4,273
ARED personnel	1,238,417	2,135
Non-ARED personnel	1,240,000	2,138
Non-personnel	15,402,471	26,556
Travel	15,381,557	26,520
Equipment/Materials	20,913	36
Facility	0	0
Overhead	1,986,765	3,425
Activity total	19,867,653	34,255
Direct instruction		
Personnel	92,867,258	160,116
ARED personnel	56,567,258	97,530
Non-ARED personnel	36,300,000	62,586
Non-personnel	47,037,704	81,099
Travel	0	0
Equipment/Materials	47,037,704	81,099
Facility	0	0
Overhead	15,544,996	26,802
Activity total	155,449,958	268,017
Pedagogical days		
Personnel	2,061,988	3,555
ARED personnel	981,988	1,693
Non-ARED personnel	1,080,000	1,862
Non-personnel	14,780,177	25,483

Resource category	Cost (FCFA)	Cost (USD)
Travel	14,759,264	25,447
Equipment/Materials	20,913	36
Facility	0	0
Overhead	1,871,352	3,226
Activity total	18,713,517	32,265
Management, planning, and evaluation		
Personnel	30,657,927	52,858
ARED personnel	29,217,927	50,376
Non-ARED personnel	1,440,000	2,483
Non-personnel	20,741,202	35,761
Travel	20,564,279	35,456
Equipment/materials	151,923	262
Facility	25,000	43
Overhead	5,711,014	9,847
Activity total	57,110,144	98,466
TOTAL	**251,141,271**	**433,002**

Appendix D. Robustness Checks

Exhibit D131. Selection into *Ndaw Wune* Literacy Groups at Baseline



Note. Dalberg used the same selection test as ARED uses in the *Ndaw Wune* program to determine eligible students. Literacy Group 1 students read fewer than 10 letters. Literacy Group 2 students read 10 letters but fewer than 10 syllables. Literacy Group 3 students read 10 letters and 10 syllables. N = 561 comparison students and 405 treatment students.
**p* < .10. ***p* < .05. ****p* < .01.

Exhibit D232. Baseline Equivalence of Missing Versus Matched Students

Outcome	Comparison missing	Comparison matched	Difference	Treatment missing	Treatment matched	Difference
Girl	65%	47%	−19%***	51%	50%	−2%
Age	9.55 years	9.49 years	−0.06 years	9.52 years	9.69 years	0.17 years
Grade CE1	97%	95%	−2%	93%	97%	4%
Attended a preschool	64%	56%	−7%	62%	59%	−3%
LOI: Wolof	70%	65%	−5%	0.43	0.3	−0.13*
LOI: Serer	2%	3%	1%	0.35	0.35	0.01
LOI: Pulaar	28%	32%	4%	0.23	0.35	0.12**
Eaten food today	81%	83%	3%	81%	82%	1%
Received tutoring outside of school	70%	78%	7%	88%	88%	0%
During the last full week of school, how many days did you miss?	1.56 days	1.56 days	0 days	1.57 days	1.31 days	−0.26 days**

Outcome	Comparison missing	Comparison matched	Difference	Treatment missing	Treatment matched	Difference
Listening comprehension (out of 6)	3.25	3.11	−0.14	3.36	3.42	0.06
Phonological awareness (out of 10)	7.83	7.53	−0.29	8.68	8.39	−0.28
Letter sound knowledge (out of 100)	72.04	76.15	4.1	69.25	71.25	2
Decoding invented words (out of 50)	6.13	6.78	0.65	5.58	6.12	0.54
Decoding familiar words (out of 50)	5.39	6.29	0.9	5.99	6.99	1
Oral reading fluency (out of 50)	4.09	5.53	1.45*	5.35	5.64	0.29
Reading comprehension (out of 6)	0.48	0.56	0.08	0.56	0.72	0.16
Number ID (out of 24)	17.64	18.18	0.54	17.82	18.36	0.54
Counting (out of 6)	5.18	5.09	−0.09	5.37	5.32	−0.05
Digit discrimination (out of 12)	9.83	9.71	−0.11	10.37	10.4	0.04
Number sense (out of 12)	4.43	4.43	−0.01	4.69	4.41	−0.28
Addition (out of 12)	6.03	6.25	0.22	6.71	7.11	0.4
Subtraction (out of 12)	4.4	4.86	0.46	4.78	5.23	0.45

Note. LOI = language of instruction. Oral reading fluency was the only timed test at 60 seconds.
*p < .10. **p < .05. ***p < .01.

Exhibit D333. Baseline Equivalence for Matched Sample

Outcome	Comparison group	Treatment group	Difference
Girl	47%	50%	3%*
Age	9.49 years	9.69 years	0.19 years
Grade CE1	95%	97%	2%
Attended a preschool	56%	59%	3%

Outcome	Comparison group	Treatment group	Difference
LOI: Wolof	65%	30%	−35%***
LOI: Serer	3%	35%	32%***
LOI: Pulaar	32%	35%	3%
Eaten food today	83%	82%	−1%
Received tutoring outside of school	78%	88%	11%***
During the last full week of school, how many days did you miss?	1.56 days	1.31 days	−0.24 days***
Listening comprehension (out of 6)	3.11	3.42	0.31
Phonological awareness (out of 10)	7.53	8.39	0.86*
Letter sound knowledge (out of 100)	76.15	71.25	−4.9*
Decoding invented words (out of 50)	6.78	6.12	−0.65
Decoding familiar words (out of 50)	6.29	6.99	0.70
Oral reading fluency (out of 50)	5.53	5.64	0.11
Reading comprehension (out of 6)	0.56	0.72	0.16
Number ID (out of 24)	18.18	18.36	0.18
Counting (out of 6)	5.09	5.32	0.23**
Digit discrimination (out of 12)	9.71	10.4	0.69**
Number sense (out of 12)	4.43	4.41	−0.02
Addition (out of 12)	6.25	7.11	0.86*
Subtraction (out of 12)	4.86	5.23	0.37

Note. LOI = language of instruction. Oral reading fluency was the only timed test at 60 seconds.
*p < .10. **p < .05. ***p < .01.

Exhibit D434. Baseline Equivalence Without Serer Students

Outcome	Comparison group	Treatment group	Difference
Girl	50%	51%	0%
Age	9.51 years	9.59 years	0.07 years
Grade CE1	95%	94%	−1%
Attended a preschool	59%	58%	−1%
LOI: Wolof	68%	51%	−17%
LOI: Serer	0%	0%	0%
LOI: Pulaar	32%	49%	17%
Eaten food today	83%	77%	−6%

Outcome	Comparison group	Treatment group	Difference
Received tutoring outside of school	76%	88%	12%***
During the last full week of school, how many days did you miss?	1.55 days	1.38 days	−0.17 days**
Listening comprehension (out of 6)	3.17	3.71	0.54**
Phonological awareness (out of 10)	7.58	7.69	0.11
Letter sound knowledge (out of 100)	75.74	74.98	−0.75
Decoding invented words (out of 50)	6.78	7.21	0.43
Decoding familiar words (out of 50)	6.16	6.36	0.20
Oral reading fluency (out of 50)	5.37	6.61	1.25
Reading comprehension (out of 6)	0.56	0.95	0.39
Number ID (out of 24)	18.19	18.67	0.47
Counting (out of 6)	5.12	5.4	0.28**
Digit discrimination (out of 12)	9.75	10.37	0.62*
Number sense (out of 12)	4.49	5.18	0.69
Addition (out of 12)	6.28	7.17	0.89*
Subtraction (out of 12)	4.81	5.24	0.43

Note. LOI = language of instruction. Oral reading fluency was the only timed test at 60 seconds.
**p* < .10. ***p* < .05. ****p* < .01.

Exhibit D535. Random Effects Model

Outcome	Comparison baseline	Comparison endline	Treatment baseline	Treatment endline	Standardized treatment effect	*N*
Listening comprehension (out of 6)	3.14	4.22	3.40	4.41	−0.06 (0.10)	2,227
Phonological awareness (out of 10)	7.59	7.35	8.46	8.74	0.15 (0.13)	2,227
Decoding construct	75.31	77.69	70.75	83.52	0.46*** (0.10)	2,227
• Letter sound knowledge (out of 100)	n/a	n/a	n/a	n/a	0.42*** (0.08)	2,227
• Invented word decoding (out of 50)	6.64	9.94	5.99	11.85	0.33*** (0.08)	2,227

• Familiar word decoding (out of 50)	6.11	9.00	6.74	12.87	0.48*** (0.10)	2,227
Oral reading fluency (out of 50)	5.24	9.88	5.57	13.12	0.32*** (0.11)	2,227
Reading comprehension (out of 6)	0.54	0.97	0.68	1.35	0.14* (0.08)	2,191
Number ID (out of 24)	18.07	20.34	18.23	21.00	0.11 (0.07)	2,227
Counting (out of 6)	5.11	5.51	5.33	5.69	−0.03 (0.11)	2,227
Digit discrimination (out of 12)	9.74	10.85	10.39	11.22	−0.10 (0.08)	2,227
Number sense (out of 12)	4.43	5.25	4.48	5.84	0.18** (0.08)	2,227
Addition (out of 12)	6.21	8.27	7.01	9.28	0.06 (0.08)	2,227
Subtraction (out of 12)	4.76	6.22	5.12	7.26	0.20** (0.09)	2,227

Note. Controls include gender, grade, whether the student received outside tutoring, and the language in which the survey was administered. Oral reading fluency was the only timed test at 60 seconds.
*$p < .10$. **$p < .05$. ***$p < .01$.

Exhibit D636. Full Sample With Additional Control Variables

Outcome	Comparison baseline	Comparison endline	Treatment baseline	Treatment endline	Standardized treatment effect	*N*
Listening comprehension (out of 6)	3.14	4.22	3.40	4.41	−0.04 (0.10)	2,192
Phonological awareness (out of 10)	7.59	7.35	8.46	8.74	0.11 (0.13)	2,192
Decoding construct	n/a	n/a	n/a	n/a	0.41*** (0.08)	2,192
• Letter sound knowledge (out of 100)	75.31	77.69	70.75	83.52	0.41*** (0.09)	2,192
• Invented word	6.64	9.94	5.99	11.85	0.34***	2,192

Outcome	Comparison baseline	Comparison endline	Treatment baseline	Treatment endline	Standardized treatment effect	*N*
decoding (out of 50)					(0.09)	
• Familiar word decoding (out of 50)	6.11	9.00	6.74	12.87	0.50*** (0.10)	2,192
Oral reading fluency (out of 50)	5.24	9.88	5.57	13.12	0.35*** (0.11)	2,192
Reading comprehension (out of 6)	0.54	0.97	0.68	1.35	0.16** (0.08)	2,156
Number ID (out of 24)	18.07	20.34	18.23	21.00	0.09 (0.07)	2,192
Counting (out of 6)	5.11	5.51	5.33	5.69	−0.01 (0.11)	2,192
Digit discrimination (out of 12)	9.74	10.85	10.39	11.22	−0.11 (0.08)	2,192
Number sense (out of 12)	4.43	5.25	4.48	5.84	0.19** (0.08)	2,192
Addition (out of 12)	6.21	8.27	7.01	9.28	0.07 (0.09)	2,192
Subtraction (out of 12)	4.76	6.22	5.12	7.26	0.21** (0.09)	2,192

Note. Controls include female, grade, whether the student received outside tutoring, the language in which the survey was administered, the self-reported average number of school days missed the previous week, and whether the school had a wash station, and whether that wash station had soap. Oral reading fluency was the only timed test at 60 seconds.
*p < .10. **p < .05. ***p < .01.

Exhibit D737. Matched Sample

Outcome	Comparison baseline	Comparison endline	Treatment baseline	Treatment endline	Standardized treatment effect	*N*
Listening comprehension (out of 6)	3.11	4.22	3.42	4.35	−0.08 (0.11)	1,716

Outcome	Comparison baseline	Comparison endline	Treatment baseline	Treatment endline	Standardized treatment effect	*N*
Phonological awareness (out of 10)	7.53	7.22	8.39	8.79	0.21 (0.13)	1,716
Decoding construct	n/a	n/a	n/a	n/a	0.41*** (0.08)	1,716
• Letter sound knowledge (out of 100)	76.15	76.48	71.25	83.55	0.52*** (0.10)	1,716
• Invented word decoding (out of 50)	6.78	9.7	6.12	11.13	0.27*** (0.09)	1,716
• Familiar word decoding (out of 50)	6.29	8.89	6.99	12.51	0.43*** (0.11)	1,716
Oral reading fluency (out of 50)	5.53	9.59	5.64	12.3	0.3** (0.11)	1,716
Reading comprehension (out of 6)	0.56	0.96	0.72	1.20	0.05 (0.08)	1,694
Number ID (out of 24)	18.18	20.28	18.36	20.93	0.11* (0.07)	1,716
Counting (out of 6)	5.09	5.53	5.32	5.71	−0.01 (0.11)	1,716
Digit discrimination (out of 12)	9.71	10.85	10.40	11.19	−0.12 (0.08)	1,716
Number sense (out of 12)	4.43	5.23	4.41	5.61	0.16* (0.09)	1,716
Addition (out of 12)	6.25	8.20	7.11	9.17	0.06 (0.09)	1,716
Subtraction (out of 12)	4.86	6.25	5.23	7.17	0.16** (0.09)	1,716

Note. Controls include gender, grade, whether the student received outside tutoring, and the language in which the survey was administered. Oral reading fluency was the only timed test at 60 seconds.
**p* < .10. ***p* < .05. ****p* < .01.%

Exhibit D838. Full Sample without Serer Students

Outcome	Comparison baseline	Comparison endline	Treatment baseline	Treatment endline	Standardized treatment effect	*N*
Listening comprehension (out of 6)	3.17	4.25	3.71	4.42	−0.21** (0.10)	1,813
Phonological awareness (out of 10)	7.58	7.4	7.69	8.6	0.31** (0.16)	1,813
Decoding construct	n/a	n/a	n/a	n/a	0.42*** (0.09)	1,813
• Letter sound knowledge (out of 100)	75.74	78.29	74.98	86	0.36*** (0.10)	1,813
• Invented word decoding (out of 50)	6.78	10.1	7.21	13.54	0.38*** (0.11)	1,813
• Familiar word decoding (out of 50)	6.16	9.04	6.36	12.73	0.50*** (0.13)	1,813
Oral reading fluency (out of 50)	5.37	10.05	6.61	14.63	0.36** (0.15)	1,813
Reading comprehension (out of 6)	0.56	0.98	0.95	1.56	0.10 (0.10)	1,783
Number ID (out of 24)	18.19	20.42	18.67	20.92	0.01 (0.08)	1,813
Counting (out of 6)	5.12	5.51	5.4	5.62	−0.14 (0.11)	1,813
Digit discrimination (out of 12)	9.75	10.85	10.37	11.18	−0.10 (0.09)	1,813
Number sense (out of 12)	4.49	5.32	5.18	6.27	0.10 (0.10)	1,813
Addition (out of 12)	6.28	8.33	7.17	9.02	−0.05 (0.07)	1,813
Subtraction (out of 12)	4.81	6.26	5.24	6.94	0.06 (0.09)	1,813

Note. Controls include gender, grade, whether the student received outside tutoring, and the language in which the survey was administered. Oral reading fluency was the only timed test at 60 seconds.
*$p < .10$. **$p < .05$. ***$p < .01$.%

Appendix E. Inter-rater Reliability

To measure the reliability and level of homogeneity of enumerators' scores on children's literacy skills, 15% of the baseline and 20% of the endline sample were assessed by two different enumerators simultaneously. Long, one-way Analysis of Variance techniques, which are used to determine whether the mean of a dependent variable is the same in two or more unrelated and independent groups, were used to calculate the intra-class correlation within pairs of assessors for a measure of inter-rater reliability. Adapted from Fleiss et al. (1973), we interpreted the intra-class correlations as follows:

- Less than .40 – Poor
- Between .40 and .75 – Good or fair
- Greater than .75 – Excellent

Exhibit E1 shows the percentage of agreement between the raters, as well as IRR ratings for the project evaluation sample at baseline and endline. Overall, the IRR across the project evaluation sample was excellent for all variables. Therefore, the ANOVA test could not calculate the IRR. Given that this means the results were identical, it is a positive result.

Exhibit E139. Inter-rater Reliability by Literacy Skill Subtest for Sample

Literacy Skill Subtest	Baseline		Endline	
	IRR	Rating	IRR	Rating
Age	n/a	Excellent	n/a	Excellent
During the last full week of school, how many days did you miss?	100%	Excellent	n/a	Excellent
Eaten food today	96%	Excellent	90%	Excellent
Semantic Fluency	97%	Excellent	100%	Excellent
Phonological awareness	99%	Excellent	100%	Excellent
Letter sound knowledge	99%	Excellent	100%	Excellent
Decoding invented words	99%	Excellent	100%	Excellent
Decoding familiar words	100%	Excellent	99%	Excellent
Oral reading fluency	100%	Excellent	100%	Excellent
Reading comprehension	99%	Excellent	100%	Excellent
Number ID	n/a	Excellent	n/a	Excellent
Counting	98%	Excellent	90%	Excellent
Digit discrimination	n/a	Excellent	n/a	Excellent
Number sense	99%	Excellent	n/a	Excellent
Addition	99%	Excellent	99%	Excellent
Subtraction	98%	Excellent	98%	Excellent

Note. An IRR of n/a means that there was no variation between enumerators which means the ANOVA regression could not calculate an inter-class correlation. *N = 165 at baseline and 229 at endline.*

Appendix F. Survey Instruments

Pre-survey Questions

To be answered by the enumerator prior to the survey.

1. GPS coordinates *(Program the survey to collect GPS coordinates.)*

2. Select the region.

   a. Diourbel

   b. Kaolack

   c. Matam

   d. Saint Louis

3. Select the school name. *(Select from dropdown of preloaded names.)*

4. Is this a treatment school?

   a. Yes

   b. No

-- *Questions 5 and 6 are for endline only.* --

5. Was this student interrogated at baseline or is this a replacement student?

   a. Student from baseline

   b. Replacement

6. *(If A was selected in question 5.)* Select the name of the student. *(Select from a preloaded list of names. The programmer can filter the list to only show the students in the selected school.)*

--

7. Is this an individual or pair assessment? (For quality assurance purposes.)

   a. Individual assessment à skip to next section.

   b. Pair assessment

8. (If B is selected for question 7.) Are you the observing or talking enumerator?

   a. Observing

   b. Talking

Background Information

NOT TIMED

NO STIMULUS SHEET

Read each question exactly as it is written. If the child seems not to understand, you may rephrase the question slightly to help them understand. Also provide the child with adequate time to respond to each question before proceeding to the next.

1. What is your name? *[Note to enumerator: Please split the name into first name and last name as appropriate. If there is no first name or last name, please enter NA for that field]*

2. Child's gender:

   a. Boy

   b. Girl

   c. other/unknown

3. How old are you? Enter age in years…… *(Write '999' if age is unknown)*

4. What grade are you in now?

   a. CP

   b. CE1

5. Did you attend a learning program (i.e., preschool) before grade 1?

   a. No

   b. Yes

   c. Don't know/no answer.

6. What language(s) do you speak at home? (Select all that apply.):

   a. French

b. Wolof

c. Serer

d. Pulaar

e. Others, specify.

7. What language(s) do you speak on the playground with your friends? (Select all that apply):

    a. French

    b. Wolof

    c. Serer

    d. Pulaar

    e. Others, specify.

8. During the last full week of school, how many days did you **miss**?

    a. 0 days, I was at school every day.

    b. 1 day

    c. 2 days

    d. 3 days

    e. 4 days

    f. 5 days, I missed every day of school.

    g. Don't know/Didn't answer.

9. What was the reason/reasons that you missed school? (*Only ask if they missed any days. Do not read answer choices. Select all that apply)*

    a. Child was sick.

    b. Rain / poor weather

    c. Teacher was absent.

    d. I did not want to go

e. Could not pay for school supplies or school fees.

f. Needed to help with housework

g. Needed to help parents with their paid work/income job or worked for wages.

h. Transport issues/or lack of transport.

i. Taking care of siblings

j. Nobody supports me to go to school.

k. Other (Specify …………………….)

10. Have you eaten today?

 a. Yes

 b. No

11. *(baseline only)* Did you participate in the Ndaw Wune after school program during the last school year? *(Programmer: Make this question for CE1 students only in schools that received the programming.)*

 a. Yes

 b. No

 c. Don’t know/Didn’t answer.

(endline only) In what month did you begin attending the Ndaw Wune classes? (Programmer: Make this question only for the students in treatment schools.)

 a. December

 b. January

 c. February

 d. March

 e. April

 f. May

 g. None of the above

h. Don’t know/Didn’t answer.

11b. How often did you attend the Ndaw Wune classes?

a. Generally three days per week

b. Generally two days per week

c. Generally one day per week

d. Rarely (less than one day per week)

11c. Since the beginning of the program in December, how many Ndaw Wune classes did you miss?

a. Never missed a class

b. Missed 1 – 5 classes

c. Missed 6 – 10 classes

d. Missed more than 10 classes

12. *(baseline)* In the past year, have you participated in any tutoring outside of school? *(Enumerator: You may explain that “tutoring” refers to a teacher, family member, or another student giving extra lessons outside of school for reading or math. This would be in addition to the Ndaw Wune tutoring if the student received that.)*

(endline) Over the course of the current school year, did you receive any tutoring outside of the school day, in addition to Ndaw Wune (if applicable)? (*Enumerator: You may explain that “tutoring” refers to a teacher, family member, friend, or other student who gives additional lessons outside of the school day for reading or math.)*

a. Yes

b. No

c. Don’t know/Didn’t answer

12b. (if student responds yes) How much tutoring did you receive in the past year?

a. Less than 5 hours total

b. 5-10 hours total

c. More than 10 hours total

d. Don't know/Didn't answer

Reading Assessment

Listening comprehension

For this activity, the children should be tested in the language of instruction used at the school (Wolof, Pulaar, or Serer).

NOT TIMED

NO STIMULUS

Now I'm going to read you a story twice. After that, I'm going to ask you some questions about this story. You will listen well, and then you will answer the questions as best you can. Okay? Do you understand what I'm asking you to do? Let's start! Listen carefully:

- It's the end of the year.
- The school director prepares for the party with the teachers and students.
- The girls do the shopping. The boys bring wood and water.
- The teachers prepare the meal.
- We eat, we dance, we play. Everybody is happy.

Evaluator: Repeat the story, then pose the questions**.**

1. What time of year is the party? Correct *(end of the year)*/Incorrect/No answer
2. What does the director do? Correct *(prepares for the party)*/Incorrect/No answer
3. What are the girls doing? Correct *(shopping)*/Incorrect/No answer
4. What are the boys doing? Correct *(bringing wood and water)*/Incorrect/No answer
5. Who prepares the meal? Correct *(the teachers)*/Incorrect/No answer
6. Why is everyone happy? Correct (because they are singing, dancing, and playing; because it's fun, it's a party, etc.)/Incorrect/No answer]

Semantic Fluency

NOTE: This section is for all students.

ENUMERATOR INSTRUCTIONS:

This section should be repeated in each language that the child indicated in questions 6 and 7. Also test the child in the national language used at school (Wolof, Serer, or Pulaar) if the child did not indicate these languages.

Point the child to the image which includes a scene with multiple images and the child will try to say as many words as they can think of about the scene - or any other word that comes to their mind - in 60 seconds.

It is important that this is NOT a test of the words in the picture (i.e., a vocabulary test), but rather a test of the number and speed with which words come to a child's mind in a given language. As long as they are correct words in the language in which the test is being conducted, it is good to mark the word as correct. For example, if there is a pig in the picture, and the child says "pig" or "dog" or "farm" or any real word in that language, it should still be marked as a correct word. Note that the child is not required to name all objects visible in the scene.

Record the number of words the student says during the 60 second period. After a break, the test is repeated with the same child, using the same scene but in another language, and a new score is recorded.

PROGRAMMER NOTES: We have 2 images that we need to recycle depending on the number of languages. For example, if the child says he/she speaks:

- 1 language, then we will use Image 1.
- 2 languages, then we will use Image 1 for the first language and Image 2 for the second language.
- 3 languages, then we will use Image 1 for the first language, Image 2 for second language, and Image 1 again for the third language.
- 4 languages, then we will use Image 1 for the first language, Image 2 for the second language, Image 1 again for the third language, and Image 2 again for the fourth language
- And so on…

Image 1

Image 2

Phonological Awareness

NOTE: This section is for all students.

For this activity, the children should be tested in the language of instruction used at the school (Wolof, Pulaar, or Serer).

NOT TIMED

NO STIMULUS SHEET

STOP RULE: If the student fails to give a single correct answer among the first five words, ask him nicely to stop and tick the "Stop" box. Go to the next exercise.

ADMINISTRATION: Enumerator should click on the tablet option only when the child does not give a correct answer. If the child gives the correct answer, do not click on the tablet option.

[PROGRAMMER INSTRUCTION: INCLUDE AUTOSTOP IF THE FIRST FIVE OPTIONS ARE WRONG]

This exercise **is only done orally. Do not show the written words to the student.** The test is not timed. First give the student the examples one after the other explaining the task. After each example, ask him/her to answer on his/her own.

INSTRUCTIONS TO STUDENT:

This exercise is an oral exercise. I'll say a word to you twice, then I want you tell me the very first sound of the word you hear, okay?

For example: The word “soupe” starts with the sound “sssss”, doesn't it? What is the very first sound in the word “soupe”? "Soupe"? [Wait for the student to repeat the “sssss” sound. If he/she doesn't answer, say, “The very first sound of the word 'soup' is 'sssss'].

Let's try a few more examples:

What is the very first sound in the word " jour"? If the student answers correctly, say “Very well! The first sound in the word “jour” is “jjjj”. If the student does not respond, say, “The first sound in the word “day” is “jjjj.”

What is the very first sound in the word "chic"? If the student answers correctly, say “Very well! The first sound in the word “chic” is “ch”. If the student does not respond, say, “The first sound in the word “chic” is “ch.”

What is the very first sound in the word " poulet"? If the student answers correctly, say “Very well! The first sound in the word “poulet” is “p”. If the student does not respond, say, "The first sound in the word poulet is 'p'.

Do you understand what I'm asking you to do? Now I will read you some other words. I will read every word twice. Listen carefully, and tell me the very first sound you hear, okay?

Do not correct the student during the test. In case of no answer or hesitation on his part, after 3 seconds mark the “No Response” box and move on to the next item.

Stop rule: If the student fails to give a single correct answer among the first five words, ask him nicely to stop and tick the “Stop” box. Go to the next exercise.

1. jour	**/j/**
2. dur	**/d'/**
3. lac	**/l/**
4. sac	**/s/**
5. car	**/k'/**
6. par	**/p'/**
7. tour	**/t'/**
8. fil	**/f/**
9. balle	**/b'/**
10. vol	**/v/**

Letter Knowledge

NOTE: This section is for all students.

For this activity, the children should be tested in the language of instruction used at the school (Wolof, Pulaar, or Serer.)

STIMULUS SHEET MUST BE PROVIDED TO STUDENT.

STOP RULE: Stop the test if the student makes 10 consecutive errors.

ADMINISTRATION:

Follow the screen and mark any incorrect letter by tapping that letter on the screen – it will turn blue. Mark auto-corrections as correct by tapping the letter again - it will return to white.

Be silent unless the child hesitates for 3 seconds. If the child hesitates for 3 seconds, point to the next letter and say, **“Please go on."** Mark the ignored letter as incorrect.

If the student says the name of the letter instead of the sound, say, **“Please tell me the sound of the letter."** This warning may be given only once during the exercise. If the time ends before the student reads the last item, the screen will flash red and the stopwatch will stop. Tell the child to stop.

Mark the total number of letters read and the number of letters read incorrectly.

INSTRUCTIONS TO STUDENT:

Here is a page filled with letters of the alphabet. Please tell me the SOUND of as many letters as you know, not their NAME, but the SOUND they make. For example (point to the letter C) the sound of this letter is sometimes /k/ as in concombre or sometimes /s/ as in citron.

Let's try together! Tell me the sound of this letter (show the letter V). If the child answers correctly, say: Good! The sound of this letter is /v/. If the child does not answer correctly, say: the sound of this letter is /v/

Let's try another letter! Tell me the sound of this letter (show the letter L). If the child answers correctly, say: Good! The sound of this letter is /l/. If the child does not answer correctly, say: the sound of this letter is /l/

Did you understand?

When I say "begin", you are going to tell me the sound of the letters, as fast as you can but still being careful. You are going to start saying the sound of the letter starting here and continuing this way. (Indicate the first letter of the first line and run your finger in the direction of the reading of the first line.)

If you get to a letter you don't know, move on to the next letter. Put your finger on the first letter. Are you ready? Begin!

E	f	O	i	ou	A	z	é	ch	Q
b	on	s	N	T	i	G	m	L	an
w	g	ou	O	M	L	p	T	j	c
V	a	R	K	b	u	s	f	é	J
s	c	an	L	e	D	a	Y	f	H
i	U	p	s	P	M	n	v	oi	T
Z	e	g	un	v	in	an	F	d	o
d	b	A	é	r	m	o	on	T	C
R	q	B	L	ou	e	p	n	i	a
gn	ch	V	E	x	d	m	U	ç	oi

Decoding

NOTE: This section is for all students.

For this activity, the students should be tested in the language of instruction used at the school (Wolof, Pulaar, or Serer.)

Reading Familiar Words

STIMULUS SHEET MUST BE PROVIDED TO STUDENT.

STOP RULE: If the student fails to give a single correct answer for five consecutive items, ask him nicely to stop and tick the “Stop” box. Go to the next exercise.

ADMINISTRATION:

Follow the screen and mark any incorrect word by tapping that word on the screen – it will turn blue. Mark auto-corrections as correct by tapping the word again - it will return to white.

Be silent unless the child hesitates for 5 seconds. If the child hesitates for 5 seconds, point to the next word and say, **“Please go on."** Mark the ignored word as incorrect.

[PROGRAMMER INSTRUCTION: INCLUDE OTHER INSTRUCTIONS AS PER ADMINISTRATION NOTES ABOVE. MARK TOTAL WORDS READ, NUMBER OF WORDS READ INCORRECTLY AND TOTAL TIME TAKEN TO READ THE WORDS.]

Here is a page with words you are going to read. Here are 3 examples. This first word [Indicate the word “ta” with the finger] reads “ta”. Can you read that first word? [After she answers, or after 5 seconds if there is no answer, show her how to do it.] And this word? [indicate the word “par” with your finger]. Can you read me this word? [After she answers, or after 5 seconds if there is no answer, show her how to do it.] And this word? [indicate the word “lune” with your finger]. Can you read me this word? [After she answers, or after 5 seconds if there is no answer, show her how to do it.] Okay?

Do you understand what I'm asking you to do? When I say "Begin", you will read the words from left to right, line by line. At the end of a line, you will move on to the next line. Try to read quickly and correctly. Are you ready? Begin.

Start the timer when the student tries the first word (“you”), by pressing the button.

tu	ou	il	ma	vol
sa	ami	or	car	lire
sol	sage	peur	bébé	papa
carte	blé	cri	fleur	vache
sur	vole	chaise	bleu	peau
mil	clé	mur	monde	table
fin	posé	date	kilo	tour
ronde	faire	pré	porter	abri
été	rougir	beau	moto	pain
mal	vélo	douze	vide	bol

Reading Invented Words

STIMULUS SHEET MUST BE PROVIDED TO STUDENT.

STOP RULE: If the student fails to give a single correct answer for five consecutive items, ask him nicely to stop and tick the “Stop” box. Go to the next exercise.

ADMINISTRATION:

Follow the screen and mark any incorrect word by tapping that word on the screen – it will turn blue. Mark auto-corrections as correct by tapping the word again - it will return to white.

Be silent unless the child hesitates for 5 seconds. If the child hesitates for 5 seconds, point to the next word and say, **“Please go on."** Mark the ignored word as incorrect.

[PROGRAMMER INSTRUCTION: INCLUDE AUTOSTOP IF THE FIRST FIVE OPTIONS ARE WRONG. INCLUDE OTHER INSTRUCTIONS AS PER ADMINISTRATION NOTES ABOVE. MARK TOTAL WORDS READ AND THE NUMBER OF WORDS READ INCORRECTLY.]

Here are some words you may never have seen. But I would like you to try to read them. For example, this first word [Indicate the word "bi" with your finger] reads "bi". Can you read that first word? [After she answers, or after 5 seconds if there is no answer, show her how to do it.] And this word? [indicate the word “sep” with your finger]. Can you read me this word? [After he/she answers, or after 5 seconds if there is no answer, show him/her how to do it.] And this word? [indicate the word “lar” with your finger]. Can you read me this word? [After she answers, or after 5 seconds if there is no answer, show her how to do it.] Okay?

Do you understand what I'm asking you to do? When I say "Begin", you will read the words from left to right, line by line. At the end of a line, you will move on to the next line. Try to read quickly and correctly. Are you ready)? Begin.

Start the timer when the student tries the first word (“ja”), by pressing the button

ja	tal	zi	ol	vaf
sar	cla	vor	ciko	ul
bige	plovi	neul	bilba	ima
tipa	blu	osi	toche	flir
saré	rané	nur	pro	duse
mouli	clo	chane	doupé	bape
til	zopé	taindé	nube	doul
donré	raite	dreu	lorpe	ibrau
oti	nogir	neau	moudir	bir
bair	lépa	zode	fipe	nour

Oral Reading Fluency

NOTE: This section is for students who obtained a score greater than 5 in the preceding section.

For this activity, the students should be tested in the language of instruction used at the school (Wolof, Pulaar, or Serer).

TIMED: 60 seconds

STIMULUS SHEET/SHOWCARD MUST BE PROVIDED TO STUDENT.

STOP RULE: If the student does not read a single word among the words required to answer the first question, ask him nicely to stop and tick the “Stop” box. Go to the next exercise.

ADMINISTRATION: Start the timer when the child reads the first word.

Follow the screen and mark any incorrect word by tapping that word on the screen – it will turn blue. Mark auto-corrections as correct by tapping the word again - it will return to white.

Be silent unless the child hesitates for 5 seconds. If the child hesitates for 5 seconds, point to the next word and say, "**Please go on**." Mark the ignored word as incorrect.

If the time ends before the student reads the last word, the screen will flash red and the stopwatch will stop. Tell the child to stop. Check the final word read on the screen so that a red bracket appears. Then press "**Next**".

If the child reaches the last item before the screen is red/timer is up, stop the timer as soon as the child reads the last word. Tap the last word to have the red bracket appear. Then press "**Next**".

[PROGRAMMER INSTRUCTION: INCLUDE AUTOSTOP IF ALL THE WORDS REQUIRED TO ANSWER THE FIRST QUESTION ARE READ WRONG. INCLUDE OTHER INSTRUCTIONS AS PER ADMINISTRATION NOTES ABOVE. MARK TOTAL WORDS READ, NUMBER OF WORDS READ INCORRECTLY AND TIME TAKEN.]

PLEASE IMPLEMENT THE FOLLOWING LOGIC FOR THE QUESTIONS TO SHOW UP FROM THE NEXT SECTION:

- If the child is able to read only the first line (up to or before "s'amuse"), show only the first question ie What day do they have fun?
- If the child is able to read only the first two lines (up to or before "frère"), show only the first two questions
- If the child is able to read only the first three lines (up to or before "un ballon"), show only the first three questions
- If the child is able to read only the first four lines (up to or before "au ballon"), show only the first four questions
- If the child is able to read the first five lines (up to or before "contents"), show all the questions

Now I would like you to read aloud the story of a child. Try to read quickly and correctly; then I'm going to ask you a few questions. Start here when I tell you. (Put the showcard in front of the student. Point to the first line (the title). Are you ready?

Begin. [Start the timer by pressing the START / STOP button]

Le jeu de ballon / C’est samedi. Ali s’amuse	8
dans la cour de l’école avec son petit frère.	17
Son camarade Sidi arrive avec un ballon.	24
Ali, son frère et Sidi jouent au ballon	32
toute la matinée. Les amis ont bien joué. Ils sont contents.	43

Reading Comprehension

NOTE: This section is students reading at the rate of 10 words per minute or more in the previous section.

For this activity, the students should be tested in the language of instruction used at school (Wolof, Pulaar, or Serer).

NOT TIMED

STIMULUS FROM THE PREVIOUS ACTIVITY

[PROGRAMMER INSTRUCTION: PROGRAM QUESTIONS TO APPEAR DEPENDING ON HOW FAR THE CHILD READ THE ORAL READING PASSAGE IN THE PREVIOUS SECTION.]

When the student has finished reading the section above, remove the text from their possession and ask the first question below. If the student gives no answer after 10 seconds, repeat the question, and give the child another 5 seconds to answer. If there is still no answer, move on to the next question. Ask them questions that match lines of text up to the line where the square bracket ( ]), i.e., up to where the student stopped reading.

Record the student's responses in the "Student Responses" space: Put a cross in the box that corresponds to his answer in relation to each question.

- “Correct”: The student gives a correct answer or gave an answer incorrect but later corrected itself (self-correction). The correct answers can be provided in French or in the national language. Answers with a meaning similar to those indicated should be marked as correct.
- “Incorrect”: The student gives an incorrect answer or says I don’t know.
- “No answer”: The student does not give an answer.

Now you are going to answer some questions about the story.

- What day do they have fun? [Saturday] Correct/Incorrect/No answer

- Where are Ali and his little brother? [in the court; at school, etc.] Correct/Incorrect/No answer
- What does Sidi bring with him? [a ball] Correct/Incorrect/No answer
- Who plays ball? [Ali; Sidi; his brother; Ali and Sidi; etc.] Correct/Incorrect/No answer
- How are the children at the end of Game? [happy] Correct/Incorrect/No answer
- Why are they happy? [they have played well; etc.] Correct/Incorrect/No answer

Math Assessment

For the math assessment, enumerators should explain the activities using the language the child is most familiar with, according to the semantic fluency results. Responses in any language are acceptable.

Number Identification

LEARNING OUTCOME	Identifies numbers up to 999
MATERIALS REQUIRED	Number grid

Field Investigator Script: Let's do a number activity.

Here are some numbers. I want you to point to each number and tell me what the number is.Are you ready? Okay. Start from here.

Prompt A: What number is this?

(Note: If a child pause and stops on a number for 15 Seconds, then point to the next number andsay- what number is this?)

<u>Stop rules</u>: If child gives fewer than 4 correct responses between Rows 1 and 2, assign a 0 and move to the next activity.

If the child progresses to Rows 3 and 4, but gives fewer than 4 correct between those rows, assign a 1 and move to the next activity.

Row 1	6	8	4	2
Row 2	1	9	7	10
Row 3	1 2	1 4	1 9	20

Row 4	26	32	60	99
Row 5	100	400	900	106
Row 6	347	214	839	686

Expected Response: The child reads aloud the number as per number name convention in languageof instruction or home language.

✍ **Assessment Instruction**

Programmer / Enumerator: Please capture how many numbers from each row the student answers correctly. Include the language/languages the child uses to identify numbers.

Language(s) of response: ____________________

Assessment Rubric

Rows 1-2: Ability to identify numbers up to 10

Rows 3-4: Ability to identify numbers up to 99

Rows 5-6: Ability to identify numbers up to 999

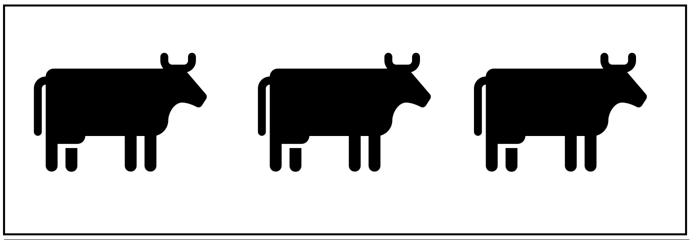
Counting

LEARNING OUTCOME	Counts numbers up to 15
MATERIALS REQUIRED	Images

Field investigator script : Let's do another activity. I will show you some pictures and I would like you to count the number of things you see in each.

Here is an example.

How many cows do you see in the picture ?



(If the child says 3) Good job, there are three cows in the picture. Let’s try some others.

(If the child says another number or does not respond, point to each cow and count.) There are three cows in this picture. Let’s try some others.

(Note: if the child stops on a problem for 30 seconds, continue to the next.)

Images

Image 1 How many girafes do you see here?

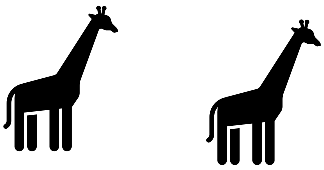


Image 2 How many birds do you see here?

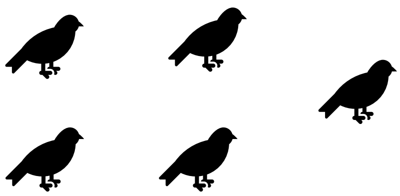


Image 3 How many pencils do you see here?

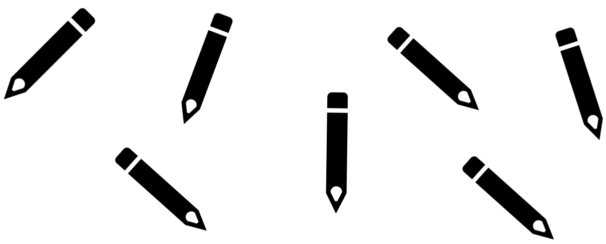


Image 4

How many cars do you see here?

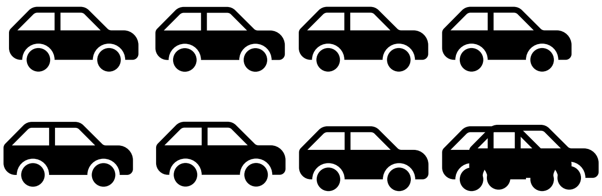


Image 5

How many trees do you see here?

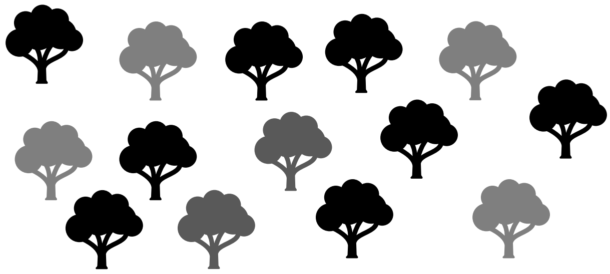
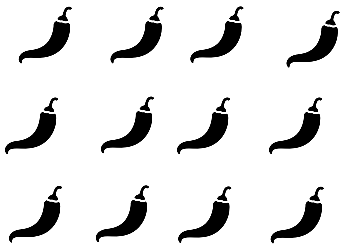


Image 6

How many peppers do you see here ?



Expected responses :

Image 1	2
Image 2	5
Image 3	7
Image 4	9
Image 5	14
Image 6	12

Assessment Instruction

Programmer/Enumerator : Note the number of correct responses.

Assessment Rubric

Images 1-2 : Ability to count quantities up to 5

Images 3-4 : Ability to count quantities up to 10

Images 5-6 : Ability to count quantities up to 15

Number discrimination

LEARNING OUTCOME	Compares numbers up to 999
MATERIALS REQUIRED	Number discrimination grid

Field investigator script : Let's do one more activity. In this activity also, I will show you some numbers.

Here's an example:

I want you to tell which number is bigger. ***(Write the numbers on a sheet and show the numbers to the child)***

6 4

(if child points to 6) That's correct, 6 is bigger. Let's continue.

(if child points to 4 or does not respond; point to 6) 6 is bigger. Let's try again.

Now, look at these numbers. Tell me which number is bigger. ***(Write the numbers on a sheet andshow the numbers to the child)***

13 19

(if child points to 19) That's correct, 19 is bigger.

(if child points to 13 or does not respond; point to 19) 19 is bigger.

Now, you will try it on your own. Look at these numbers. Tell me which number is bigger. ***(Show the numbers to the child. For the remaining numbers, do NOT indicate if the child was right or wrong or give the answer. Simply acknowledge their response and continue to the next question.)***

(Note: If a child pauses on a pair of number for 15 Seconds, then point to the nextnumber combination and say- which number is bigger?)

<u>Stop rules</u>: If child gives fewer than 2 correct responses between Rows 1 and 2, assign a 0 and move to the next activity.

If the child progresses to Rows 3 and 4, but gives fewer than 2 correct between those rows, assign a 1 and move to the next activity.

About the American Institutes for Research®

Established in 1946, the American Institutes for Research® (AIR®) is a nonpartisan, not-for-profit institution that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of education, health, and the workforce. AIR's work is driven by its mission to generate and use rigorous evidence that contributes to a better, more equitable world. With headquarters in Arlington, Virginia, AIR has offices across the U.S. and abroad. For more information, visit AIR.ORG.

AIR® Headquarters
1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
+1.202.403.5000 | **AIR.ORG**

28911_09/25